2012

Use of Exploratory Data-Mining Techniques to Analyze Associations between Bone-Mineral Density and Relevant Clinical Parameters of Gaucher Disease

Tingting Fu
Wright State University

Use of Exploratory Data-Mining Techniques to Analyze Associations between Bone-Mineral Density and Relevant Clinical Parameters of Gaucher Disease

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Engineering

By

Tingting Fu

B.S., Dalian Jiaotong University, China, 2005

2012

Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

July 12, 2012

I HEREBY RECOMMEND THAT THAT THE THESIS PREPARED UNDER MY SUPREVISION BY Tingting Fu ENTITLED Use of Exploratory Data-Mining Techniques to Analyze Associations between Bone-Mineral Density and Relevant Clinical Parameters of Gaucher Disease BE ACCEPTED IN PARTIAL FULFILIMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science in Engineering.

_____

Yan Liu, Ph.D.
Thesis Director

_____

Thomas Hangartner, Ph.D., Chair
Department of Biomedical, Industrial and Human Factors Engineering
College of Engineering and Computer Science

Committee on
Final Examination

_____

Yan Liu, Ph.D.

_____

Pratik Parikh, Ph.D.

_____

Thomas Hangartner, Ph.D.

_____

Andrew Hsu, Ph.D.
Dean, Graduate School

Abstract

Fu, Tingting. M.S.E.. Department of Biomedical, Industrial and Human Factors Engineering, Wright State University, 2012. Use of Exploratory Data-Mining Techniques to Analyze Associations between Bone-Mineral Density and Relevant Clinical Parameters of Gaucher Disease.

Gaucher disease (GD) is a monogenic disorder with autosomal recessive inheritance, which results from an acid lysosomal hydrolase, the beta-glucocerebrosidase deficiency. Clinical manifestations of the disease include anemia, thrombocytopenia, hepatosplenomegaly, and skeletal complications. Enzyme replacement therapy (ERT) has been used to treat type 1 GD for more than a decade, and many patients have shown remarkable clinical responses to the treatment, with normalization of blood counts, reduction in liver and spleen size, and improvement in bone symptoms. Many researchers have tried to study the effectiveness of ERT, but previous research has been mainly based on some predetermined hypotheses and traditional analysis methods, which assumed some statistical distributions of the underlying data. In addition, studies have suggested significant individual differences in patients' bone mineral density (BMD) responses to ERT.

In this project, we used non-parametric regression tree methods to analyze the BMD data of patients with type 1 GD, in combination with other potentially relevant parameters, including patients' demographics, hematological, visceral, and bone manifestations, to define a parameter subspace that explains the patients' BMD response. Models have been derived for the patient's initial dual-energy X-

ray absorptiometry (DXA) Z-score, the rate of change of the patient's DXA Z-scores from his/her first infusion to the current DXA assessment visit, and the rate of change of the patient's DXA Z-scores between two consecutive DXA assessment visits. Modeling results suggest that the patient's initial DXA Z-score is affected by his/her region, treatment with bisphosphonates, gender, and the period between the patient's first infusion and first DXA visit date. The rate of change of the patient's DXA Z-scores from his/her first infusion to the current DXA assessment visit is mostly related to the patient's' region, initial DXA Z-score, and ethnicity. In addition, the most predictive covariate of the rate of change of the patient's DXA Z-scores between two consecutive DXA assessment visits is the patients' immediately previous DXA Z-score.

TABLE OF CONTENTS

I. Introduction 1

1.1 What Is Gaucher Disease? 1

1.2 Enzyme Replacement Therapy and Its History 1

1.3 Limitations of Previous Research 2

1.4 Research Objective 4

1.5 Data Source of the Study 4

1.6 Contribution of the Research 5

1.7 Organization of the Thesis 5

II. Literature Review 6

2.1 Responses of GD-Related Bone Disease to Enzyme Replacement Therapy 6

2.2 Generalized Linear Mixed Model 13

2.3 Data Mining 15

2.4 Missing Data 19

2.5 Methods of Dealing with Missing Data 20

2.5.1 Maximum likelihood (ML) Method of Handing Missing Data 21

2.5.2 Bayesian Method of Treating Missing Data 23

2.5.3 Multiple Imputations 24

III. Data Preprocessing 25

3.1 Summary of Variables in the Raw Data 26

3.2 Imputation of Missing Data 29

3.3 New Variable Construction 30

3.4 Distributions of Selected Variables 31

IV. Model Development 39

4.1 Regression Tree Method 39

4.1.1 RPART Package 40

4.1.2 PARTY Package 41

4.2 Modeling Results of SPINEZ_FIRST 43

4.2.1 Models of SPINEZ_FIRST by Using PARTY package 45

4.2.2 Models of SPINEZ_FIRST Using the RPART Package 46

4.3 Modeling Results of FIRST_RATE_SPINEZ 53
4.3.1 Models of FIRST_RATE_SPINEZ Using PARTY Package 54
4.3.2 Models of FIRST_RATE_SPINEZ Using RPART Package 57
4.4 Modeling Results of RATE_SPINEZ 60
4.4.1 Models of RATE_SPINEZ Developed Using the PARTY Package 61
4.4.2 Models of RATE_SPINEZ Developed Using the RPART Package 62
4.5 Summary of Key Findings 65
4.5.1 Findings of SPINEZ_FIRST 66
4.5.2 Models of FIRST_RATE_SPINEZ 66
4.5.3 Models of RATE_SPINEZ 67
V. Discussion and Conclusions 67
List of References: 70
APPENDIX. Models of SPINEZ_FIRST 75

List of Figures

Figure **Page**

Figure 1. Histogram of AGEINF 32
Figure 2. Histogram of BISPHOS 32
Figure 3. Histogram of SEX 32
Figure 4. Histogram of REGION 32
Figure 5(a). Histogram of ETHGRP before Imputation 33
Figure 5(b). Histogram of ETHGRP after the 1st Imputation 34
Figure 5(c). Histogram of ETHGRP after the 2nd Imputation 34
Figure 6(a). Histogram of DOSE3Y before Imputation 34
Figure 6(b). Histogram of DOSE3Y after Imputation 34
Figure 7(a). Histogram of BMIB before Imputation 35
Figure 7(b). Histogram of BMIB after 1st Imputation 36
Figure 7(c). Histogram of BMIB after 2nd Imputation 36
Figure 8. Histogram of BIWEEK_FIRST_VISIT 37
Figure 9. Histogram of BIWEEK_FOLLOW_VISIT 37
Figure 10. Histogram of BIWEEK_COVISIT 37
Figure 11. Histogram of SPINEZ_FIRST 38
Figure 12. Histogram of FIRST_RATE_SPINEZ 38
Figure 13. Histogram of RATE_SPINEZ 38
Figure 14. Model of SPINEZ_FIRST derived using the PARTY Package from All the Individual Complete Datasets and the Original Dataset 46
Figure 15. Model 1 of SPINEZ_FIRST Derived Using the RPART Package from a Complete Dataset 49
Figure 16. Model 2 of SPINEZ_FIRST Derived Using the RPART Package from a Complete Dataset 50
Figure 17. Model 3 of SPINEZ_FIRST Derived Using the RPART Package from a Complete Dataset 51
Figure 18. Model 4 of SPINEZ_FIRST Derived Using the RPART Package from the Original Dataset 52
Figure 19. Model of FIRST_RATE_SPINEZ Derived Using the PARTY Package from the 12 Complete Datasets and from the Original Dataset 56

Figure 20. Model of FIRST_RATE_SPINEZ Derived Using the RPART Package from All the Complete Datasets 58

Figure 21. Model of FIRST_RATE _ SPINEZ Derived Using the RPART Package from the Original Dataset 59

Figure 22. Model of RATE_SPINEZ Derived Using the PARTY Package from All the Individual Complete Datasets and from the Original Dataset 62

Figure 23. Model of RATE_SPINEZ Derived Using the RAPRT Package from All the Complete Datasets 64

Figure 24. Model of RATE_SPINEZ Derived Using the RAPRT Package from the Original Dataset 65

List of Tables

Table **Page**

Table 1. Summary of 11 Reviewed Previous Studies on the Responses of GD-Related Bone Disease to ERT 10
Table 2. Common Techniques Used to Solve Different DM Problems 16
Table 3. Variables Related to Patients' Demographics and Characteristics 27
Table 4. Variables Related to Patients' Hematological and Visceral Manifestations 28
Table 5. Variables Related to Patients' Bone Manifestations 28
Table 6. Summary of Constructed New Variables 30
Table 7. Variables Used to Build the Models of SPINEZ_FIRST 43
Table 8. The Variables Used in Building Models of FIRST_RATE_SPINEZ 53
Table 9. Variables Used to Build Models of RATE_SPINEZ 60

ACKNOWLEDGEMENTS

I am heartily thankful to my adviser, Dr. Yan Liu, whose encouragement and guidance from the initial to the final level enabled me to develop an understanding of the project.

I'm grateful to my committee members, Dr. Thomas N. Hangartner and Dr. Pratik J. Parikh for taking their invaluable time to serve on my committee and help in improving this project.

A special thank goes to my family for their encouragement in my study in the past three years.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the project.

I. Introduction

1.1 What Is Gaucher Disease?

Gaucher disease (GD) is a monogenic disorder with autosomal recessive inheritance which results from an acid lysosomal hydrolase, the beta-glucocerebrosidase deficiency. Consequently, lysosomes of the reticuloendothelial system accumulate glycolipids, creating engorged macrophages (Gaucher cells), which displace normal tissue and result in dysfunction in many organs. Clinical manifestations of the disease include anemia, thrombocytopenia, hepatosplenomegaly, skeletal complications such as bone pain (BP) and bone crisis, cortical and medullary infarctions, cortical bone thinning, medullary expansion, osteopenia, osteolysis, osteonecrosis, and pathological fractures (Sim et al.,2008).

Clinically, GD can be classified into three types based on the presence of primary central nervous system involvement: type 1 GD is non-neuronopathic which is also the most common variant; the rare type 2 GD is neuronopathic; and type 3 GD is less common than type 1 GD and characterized by severe visceromegaly and variably progressive neurologic involvement (Hans et al., 2008).

1.2 Enzyme Replacement Therapy and Its History

Enzyme replacement therapy (ERT) is a safe, efficacious treatment for type 1 GD, which has been used for approximately 15 years. It uses alglucerase and

imiglucerase which are useful for treating GD, and patients' treatment responses can be seen within months.

Before the introduction of ERT, GD patients were mainly treated with blood infusion, total or partial splenectomy, and the use of analgesics. In some cases of rapidly progressing disease, patients can be treated, but at a high risk, through bone marrow transplantation.

GD was successfully treated with ERT for the first time in the early 1990s, using a mannose-terminated enzyme from placental tissue (alglucerase) or a recombinant enzyme (imiglucerase, both manufactured by Genzyme, Cambridge, MA). Most patients showed a remarkable clinical response to the treatment, with normalization of blood counts, reduction in liver and spleen size, and improvement in bone symptoms. At the inception of ERT, however, only a few patients could afford the therapy because of its prohibitive costs. In order to treat more GD patients, the European Cerezyme Access Programme (ECAP) started ERT in 2004.

1.3 Limitations of Previous Research

Although previous studies of GD have made significant contributions to ERT research, they were based on some predetermined hypotheses and applied traditional analysis methods, which assume some statistical distributions of the underlying data. In addition, previous studies suggest an indispensable need for further research on the effectiveness of ERT on GD-related bone disease, which is the most significant cause of immobility and long-term disability for patients. Poll et al. (2002) analyzed the lumbar spine bone-mineral density (BMD) measurements

of 30 GD patients enrolled in the International Collaborative Gaucher Group (ICGG) Gaucher Registry to determine the effect of ERT. Overall, the study found significant improvement in the BMD of the patients, yet it also showed significant individual differences in the patients' response to the ERT. In particular, some patients showed rapid improvement, but others had only minor improvement in BMD even after several years of treatment. Wenstrup et al. (2007) analyzed the data of adult patients (men, 18–70 years; women, 18–50 years) enrolled in the ICGG Gaucher Registry, for whom lumbar spine BMD measurements were available to determine the effect of ERT on BMD in type 1 GD. In the study, the sex- and age-specific reference population was first translated to the standardized BMD and used to calculate each patient's the dual-energy X-ray absorptiometry (DXA) Z-scores (Hui, 1997). The major steps of the algorithm of Z-scores are: (1) subtract the mean BMD from the individual BMD; (2) divide the result by standard deviation of BMD. Their analysis results indicated that the DXA Z-scores for patients with GD who received ERT improved significantly over time, approaching the reference population. The authors also noted a significant dose–response relationship in the ERT group. Another study led by Genzyme tested BMD improvement after ERT in children, adolescents, and adults. A significant dose-response relationship was noted for each age group, especially for the younger patients. However, the study presents very slow recovery, on average, to normal Z-scores values, which is the average of BMD of a healthy subject of the same age and gender. Based on the results of these previous studies, we assume that there

may be a combination of variables that can distinguish patients who recover bone more quickly under ERT from those who recover bone more slowly.

1.4 Research Objective

The objective of this research is to use exploratory data mining techniques to analyze the BMD data of patients with GD in combination with other potentially relevant parameters, including patients' demographics, as well as hematological, visceral, and bone manifestations, to define a parameter subspace that explains the BMD response.

1.5 Data Source of the Study

Data for this study were retrieved from the ICGG Gaucher Registry, which includes patients treated with imiglucerase between the ages of 5 and 50 years. The ICGG Gaucher Registry was first established in 1991 as a voluntary observational database to track the clinical, biochemical and therapeutic characteristics of GD patients, irrespective of their disease severity and treatment status. It consists of international and regional boards of advisors, who oversee the scientific integrity of the Gaucher Registry and who guide research, publications, policy, and the protocols for the ICGG Gaucher Registry. The ICGG Gaucher Registry is composed of anonymous clinical data of 4500 GD patients, which have been submitted by over 700 physicians from 52 countries with appropriate Institutional Review Board/Ethics Committee approval.

1.6 Contribution of the Research

The significance of the study lies in the fact that it is venturing into using previously unexplored non-parametric regression methods to study the effects of ERT on treating GD. The models resulted from the study provide valuable insights into what demographic and clinical information of GD patients affect their BMD responses to the treatment of GD using Imiglucerase.

1.7 Organization of the Thesis

The roadmap of the thesis is as follows. Chapter 2 reviews previous research on the responses of GD-related bone diseases to the ERT, generalized linear mixed models (GLMMs), the state-of-the-art modeling technique used in previous GD research studies, data mining techniques, and methods of handling missing data. The methods of the research are presented in chapters 3 and 4. In particular, chapter 3 describes how the raw data retrieved from the ICGG Gaucher Registry are preprocessed, and chapter 4 reports how models are derived from the preprocessed data and the modeling results. Finally, in chapter 5, the conclusions and discussion of the research are offered.

II. Literature Review

2.1 Responses of GD-Related Bone Disease to Enzyme Replacement Therapy

Many patients with GD suffer progressive and often disabling morbidity attributable to skeletal complications, including osteopenia, lytic lesions, pathological fractures, avascular necrosis, and joint destruction (Charrow et al., 2007). Therefore, much effort has been made to study the effectiveness of ERT on GD-related bone diseases.

Elstein et al. (1998) presented a study of examining 28 patients with varying disease severity treated with low-dose imiglucerase for 6 to 24 months. The patients were divided into two groups in terms of their frequencies of treatment: once every other week or 3 times a week. The study did not find any statistically significant difference between the two groups in the hematological parameters and organomegaly. Although the study did not describe any quantitative data of bone responses, it found that all the patients in the study who had GD-related bone problems reported subjective decreases in the intensity and frequency of bone crises.

Analyzing the data of 28 pediatric patients with GD in Italy, USA and Germany, Bembi et al. (2002) showed that the lumbar BMD of most of these patients significantly increased after 2 years of ERT, and skeletal growth rates increased among the patients exhibiting growth delays. This study suggests that ERT can improve the BMD and growth rates in pediatric patients with GD.

Poll et al. (2002) presented previous studies on the long-term effects of ERT from six data sources, with a particular focus on the response of skeletal aspects. This study showed a rapid response of bone marrow to ERT. In some patients, improvement in bone marrow was detected with MRI within the first year of treatment, but this did not reach significance until 4.5 years after starting ERT. However, although the dose of ERT may be related to bone marrow response, no significant relationship was identified in the study. In addition, the study did not find a strong relationship between age, gender, splenectomy status or genotype and the response of bone marrow to therapy.

Weinreb et al. (2002) analyzed data of 1,028 patients with type 1 GD, retrieved from the Gaucher Registry, to study 2- to 5-year effects of ERT on specific manifestations of GD, including hematologic abnormalities, organomegaly, skeletal pain, and bone crises. The study concluded that the ERT can prevent progressive manifestations of GD and ameliorate GD associated anemia, thrombocytopenia, organomegaly, bone pain, and bone crises.

Tóth et al. (2003) examined 8 patients (with ages of 3-39 years) with GD who underwent ERT for 1-8 years (30-80 IU/kg/bi-weeks/months Ceredase or Cerezyme). The study found that the use of ERT in all the patients had led to marked improvements in visceral and skeletal pathology of the patients.

Fost et al. (2006) conducted a retrospective comparative cohort study at 2 large European treatment centers. A total of 106 adult patients with type 1 GD, who started ERT, were divided into 2 groups. One received ERT with an initial dose of

no more than 50 U/kg/4 weeks (AMC), and the other received at least 60 U/kg/4 weeks (HHU). After 12 months of ERT, there were no significant differences between the two groups in their increase in platelet count and hemoglobin, and decrease in liver volume. However, patients who took the higher-dose treatment showed a quicker and better recovery in GD-related bone diseases than those with the lower-dose treatment.

Weinreb et al. (2007) investigated the impact of imiglucerase treatment on health-related quality of life (HRQOL) of patients with type 1 GD and bone involvement. 32 patients with type 1 GD with skeletal manifestations, including bone pain, medullary infarctions, avascular necrosis, and lytic lesions, received biweekly imiglucerase (at 60 U/kg), and the short Form-36 Health Survey (SF-36) was administered to assess HRQOL. After 2 years of treatment, statistically significant improvements were observed for all eight SF-36 subscales.

In Wenstrup et al. (2007), the BMD data with up to 8 years of follow-up were analyzed for 160 patients who received no ERT and 342 patients treated with ERT alone. These patients were enrolled in the ICGG Gaucher Registry. The DXA Z-scores for the patients who received ERT at a dose of 60 U/kg/2 weeks were significantly lower than those of the reference population at baseline, but they improved significantly over time. In addition, the study also noted a significant dose–response relationship in the ERT group of patients. In particular, the patients who received a higher dose of ERT had faster recovery of their BMD. However, the study also found that response to treatment was slower for BMD than for hematologic and visceral aspects of GD.

Andersson et al. (2008) analyzed data from 884 children in the ICGG Gaucher Registry to determine the effects of long-term ERT with alglucerase or imiglucerase on hematologic and visceral manifestations, linear growth, and skeletal disease. These patients had significant improvement in their BMD Z-score and bone crises as well as reduction in their liver volume size and spleen size after 8 years of treatment.

With a 48-month longitudinal cohort study of 33 patients with type 1 GD in various ethnic groups, Sims et al. (2008) reported decreases in bone pain and skeletal complications and increases in the BMD of patients treated with imiglucerase. In addition, independent of the patients' genotype, hematological, and visceral status, imiglucerase was shown to be effective for the hematological and visceral manifestations of type 1 GD.

In Mistry et al. (2011) data of 889 patients (with ages ranging from 5 to50 years) retrieved from the ICGG Gaucher Registry were analyzed. The study found improvement in GD–associated diseases, including bone problems. In addition, it suggested that the improvement of BMD as a result of ERT may be greater in younger patients than in the older adult patients.

Table 1 summarizes the 11 above reviewed previous studies on the responses of GD-related bone disease to ERT, including the main characteristics of the research subjects in the studies, the statistical methods used, and their main findings.

Table 1. Summary of 11 Reviewed Studies on the Responses of GD-Related Bone Disease to ERT

Article	Key Characteristics of Research Subjects	Duration of ERT	Statistical Analysis Method	Main Findings
Elstein et al. (1998)	28 patients	6-24 months	two sample t-test and the non-parametric Mann-Whitney test	• No statistically significant difference between the two groups in the hematological parameters and organomegaly.
Bembi et al. (2002)	28 patients Both male and female	3-9 years	N/A	• Lumbar BMD of most patients significantly increased after 2 years of ERT • Skeletal growth rates increased among the patients exhibiting growth delays
Poll et al. (2002)	More than 2000 patients from 6 different data sources, Both male and female, 5-78 years old	9 months to more than 8 years	N/A	• Rapid response of bone marrow to ERT. • No strong relationship was found between dose, age, gender, splenectomy status or genotype and the response of bone marrow to therapy.
Weinreb et al. (2002)	1028 registry patients from 25 countries United States: 541 (53%)	2-5 years	mixed model repeated-measures	ERT can prevent progressive manifestations of GD and

	Western Europe: 220 (21%) Israel: 119 (12%) elsewhere: 148 (14%)		analysis of variance, two-sample t tests, chi-squared test, and Wilcoxon rank sum test	ameliorate GD associated anemia, thrombocytopenia, organomegaly, bone pain, and bone crises.
Tóth et al. (2003)	8 patients, 8-39 years old, both male and female	1-8 years	semi-quantitative method	The use of ERT in all patients led to marked improvements in visceral and skeletal pathology of patients with Gaucher disease.
Fost et al. (2006)	Totally 106 patients 49 patients, Netherlands, 21-74 years old, both male and female 57 patients, Germany,27-82 years old, both male and female	2-4 years	Mann-Whitney U test , log-rank test, and chi-square test	The patients who received higher dose of ERT had faster recovery of their BMD.
Weinreb et al. (2007)	32 patients, less than 70 years old	4 years	Wilcoxon signed-rank test	Imiglucerase treatment had a significant positive impact on health-related quality of life of type 1 GD patients with skeletal disease, including bone infarctions, lytic lesions, and avascular necrosis.
Wenstrup et al. (2007)	342 patients,(men, 18–70 years; women, 18–50 years)	8 years	linear mixed models	The DXA Z-scores for the patients who received the ERT

				at a dose of 60 U/kg/2 weeks were significantly lower than the reference population at baseline, but they improved significantly over time. In addition, the study also noted a significant dose–response relationship in the ERT group of patients. In particular, the patients who received higher dose of ERT had faster recovery of their BMD. However, the study also found that response to the treatment was slower for BMD than for hematologic and visceral aspects of GD.
Sims et al. (2008)	33 patients, 10-70 years old, both male and female; Ethnicity groups: Ashkenazi Jewish (23 patients), Non-Jewish Caucasian (6 patients), African-American/Caribbean (1 patient), Hispanic (2 patients), American Indian	up to 48 months	linear mixed model	ERT decreased the number of bone pain and skeletal complications and increased the BMD. Despite the patients' genotype, hematological, and visceral status, ERT was shown to be effective for the

	(1 patient), others (2 patients)			hematological and visceral manifestations of type 1 GD.
Andersson et al. (2008)	884 patients, 485 of which are under 6 years old, 260 are from 6 to 12 years old, 93 are from 12 to 18 years old, and others' age are unknown	8 years	linear mixed-effects model	After 8 years of ERT, most clinical parameters studied (including anemia, platelet counts, liver and spleen volumes and bone crisis) became normal or nearly normal.
Mistry et al. (2011)	889 patients, 5-50 years old, both male and female	over 10 years	non-linear mixed models	ERT resulted in amelioration of osteopenia in all age groups, with the greatest improvements in younger patients.

2.2 Generalized Linear Mixed Model

Table 1 suggests that the linear mixed model is the most popular statistical method used to study the effectiveness of ERT over time. A generalized linear mixed model (GLMM) is a parametric linear model for clustered, longitudinal, repeated-measures data, which quantifies the relationships between multiple predictor variables and a continuous response variable (West et al., 2007).

GLMMs can be used to analyze clustered data, such as patients in different hospitals, and it also can be used in longitudinal or repeated-measures studies in which subjects are measured repeatedly in different time periods, conditions, or both (West et al., 2007).

A GLMM is a model of the form y=Xβ+Zu+e, where *y* is a vector of the continuous dependent variable, X and Z are known design matrixes, β is a vector of unknown regression coefficients associated with the X matrix, and *u* is a vector of unknown random effects associated with the *Z* matrix. The vector of *u* is assumed to be independent and identically distributed (IID), N (0, D). The elements on the diagonal of the D matrix represent the variance of each random effect, and the off-diagonal elements of the D matrix are the covariance between random effects. The vector of residuals, *e*, is assumed to be IID, N (0, R), and each element of the R matrix represents the variance of a residual or covariance between two residuals. Further, the vectors of random effects and residuals are independent from each other.

Generally speaking, there are two approaches to building GLMMs: the top-down strategy and the set-up strategy, but the top-down strategy is often used in most applications (West et al., 2007). The top-down strategy involves three steps. First, we start with a model involving enough fixed effects to explain the mean structure. Next, the random effects and residuals are added to the model. Finally, we remove from the model the fixed-effect parameters which are not significant.

After building a GLMM, it is important to carry out model diagnostics to check whether the assumptions for the residuals and random effects are satisfied, as well as whether the model is sensitive to unusual observations.

Although the GLMM provides a powerful method to analyze clustered, longitudinal and repeated-measures data, it is a rather challenging tool in many cases. GLMM makes the analyzing process complex, which may lead to imprecise

results. Since the estimation of unknown parameters depends on selected model/software, it would make the process subjective.

The GLMM assumes that random effects and residuals follow multivariate normal distributions. However, it is very difficult to check whether the multivariate normality assumption is valid or not.

Moreover, because accurate techniques for estimating GLMM parameters are only available in simple cases, complex GLMMs are challenging to fit. Therefore, we cannot add too many parameters (fixed-effect and random-effect parameters) into the GLMM, which makes this method not suitable in many applications (Brooks, 2008).

2.3 Data Mining

Data mining (DM) is a result of the natural evolution of information technology (http://dataminingtools.net/wiki/introduction_to_data_mining.php). A broader view of DM considers it as a synonym for Knowledge Discovery in Databases, or KDD. A widely accepted definition of KDD was given in Fayyard, Piatetsky-Shapiro, and Smyth (1996): KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases. DM is an iterative process that involves 6 main steps: problem understanding, data understanding, data preprocessing, modeling, evaluation, and deployment (http://www.dataminingexpertsolutions.com/dm-process/).

For the first two steps, users need to determine their objectives for carrying out DM, and collect data that is related to the DM problem. Data preprocessing is an extremely important yet often neglected step in DM, whose tasks include data cleaning, data reduction and new data construction.

Modeling is the crucial step in the DM process where some selected data mining algorithms are applied to the prepared dataset to extract patterns. Table 2 summarizes some common techniques used in different DM problems. After a model is built, it is important to evaluate its performance and results with respect to predefined objectives. Finally, if all the previous steps are satisfactory and the models fulfill the project objectives, the DM results can then be deployed in the problem domain.

Table 2. Common Techniques Used to Solve Different DM Problems

DM Problems	**Common Techniques**
Data Description	Online Analytical Processing(OLAP), Attribute-Oriented Approach, Statistical Approach
Dependency Analysis	Association Rules, Correlation Analysis, Bayesian Networks, Regression Analysis
Classification	Decision Trees, Rule Induction, Bayesian Classification, Neural Networks, K-Nearest Neighbors, Case-Based Reasoning
Prediction	Regress Analysis, Regression Trees, Neural Networks, K-Nearest Neighbors
Clustering	K-Means Methods, Hierarchical Methods, Density-Based Methods, Neural Networks, Statistical Method, Visualization
Evolution	Trend Analysis, Sequential Pattern Mining, Periodicity

Analysis	Analysis

Decision tree and regression tree are two of the most popular data mining algorithms. A decision tree is a method in the form of a tree structure, which consists of decision nodes and leaf nodes. All decision nodes have splits, testing the values of some functions of their corresponding attributes. Each branch from the decision node corresponds to a distinct outcome of the test. Each leaf node has a class label attached to it. A regression tree is an extension of the decision tree algorithm to a continuous response variable. The regression result of a leaf node in a regression tree can be shown as a single prediction value (usually the sample mean of all values of the response variable belonging to the leaf node), or a simple function (usually a linear function) that relates the response variable and covariates for the cases belonging to the leaf node, or some visual representation of the values of the response variable in the leaf node. The tree path of a leaf node in a decision tree or regression tree refers to the path of the tree from its root node, the top-level node of the tree, which represents the entire dataset, to the leaf node. Each tree path can be written in "if-then" condition statements.

Although varieties of tree algorithms have been developed with different capabilities and requirements, most are variations of a core learning algorithm that employs a "greedy" top-down search through the space of possible trees. The basic procedure is as follows:

1. Start with the root node which represents the entire training data.

2. If all the training data belong to the same class (in a decision tree) or have the same values for all the predictor variables (in a regression tree), stop, and the node becomes a leaf node.

3. Otherwise, select the splitting attribute *s*, which can "best" partition the samples based on some goodness measure of splits. This attribute becomes a decision node.

4. A branch is created for each category group of *s* if *s* is a categorical variable or each interval of *s* if *s* is a continuous variable, and the samples are partitioned accordingly.

5. The partition process continues until some tree stopping criterion is satisfied, such as the improvement is not substantial enough to justify further partitioning or no predictor attribute can be further partitioned.

When building a decision tree or regression tree, pruning is an important step to obtain a right-sized tree. Generally speaking, there are two strategies of pruning (Murthy, 1998): (1) pre-pruning, which avoids creation of more sub-trees by restricting the minimum node size, as well as thresholds on impurity and some other measures; and (2) post-pruning, which creates an overfitted tree initially and then reduces the tree size based on estimated errors.

Classification and Regression Tree (CART), which was first developed in 1980s by Breiman et al. (1984), is one of the most well-known tree algorithms. It is a nonparametric technique that can select from among a large number of variables and their interactions in determining the outcome variable.

Instead of employing stopping rules, CART generates a sequence of subtrees by first growing a full-grown tree and then pruning it back until only the root node is left. Then it uses cross-validation to estimate the misclassification cost of each subtree and chooses the one with the lowest estimated cost. In particular, CART uses the Gini index to grow a decision tree and uses the sum of squared residuals to grow a regression tree. Finally, it uses cross-validation to estimate the cost-complexity measure of each sub-tree (a measure combining the error and complexity of the subtree) and chooses the one with the minimum cost-complexity. In *k*-fold cross-validation, the original data sample is randomly divided into *k* subsamples. One of the *k* subsamples is retained as the validation data for model testing and evaluation, and the remaining $(k-1)$ subsamples are used as training data for model building. The cross-validation process is then repeated *k* times (the folds), with each of the *k* subsamples used exactly once as the validation data. The results from the *k* folds then can be combined (e.g. taking the average of the *k* results) to produce a single estimation.

2.4 Missing Data

Handling missing data is an important issue in this project because many records in the original dataset of this project have missing values. This section reviews the types of missing data mechanisms and common methods of dealing with missing values.

The impact of the missing data on the results of statistical analysis depends on the mechanism that caused the data to be missing. Generally speaking, there are 3 types of missing data mechanisms (Little & Rubin, 1987): missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). MCAR means that the missing data mechanism is unrelated to the values of any variables, whether missing or observed. MCAR is a very strong assumption because it suggests that missing data values are random samples of all data values. MAR is less restrictive than MCAR; it only requires the cause of the missing data to be unrelated to the missing values but may be related to the observed values of other variables. In other words, when MAR holds, given observed data, the missing mechanism no longer depends on the unobserved data. If the pattern of data missingness is non-random and the probabilities of nonresponse depend on the missing values themselves, then the missing data are said to be MNAR.

2.5 Methods of Dealing with Missing Data

Generally speaking, methods of handling missing data can be divided into 2 broad categories: listwise deletion (discarding the records with missing values) and missing data imputation (replacing missing data with estimated values). The first one is easy but risky, because it can lose large amounts of information in the study unless the incomplete cases comprise only a small fraction of all cases (Roth & Switzer, 1995). Therefore, only imputation techniques are reviewed in this section.

Conventional imputation includes *mean imputation*, *regression imputation*, and *hot-deck imputation* methods. *Mean imputation*, where the missing values of a

particular variable are replaced by the mean of the observed values of the variable; *regression imputation*, where missing values are imputed using the prediction from a multiple regression analysis; and *hot-deck imputation*, where missing values are replaced by the corresponding values of similar cases.

Model-based imputation procedures are advanced imputation methods which have gained much attention recently. The basic idea of model-based methods is to fit statistical models from observed data and then use the models to predict missing values. We review two model-based imputation methods –*Maximum likelihood estimation* and *Bayesian imputation* in the following paragraphs.

2.5.1 Maximum likelihood (ML) Method of Handing Missing Data

Maximum-likelihood (ML) estimation is a method of estimating parameters in a statistical model, which also can be used in an incomplete dataset (Paul et al., 2003).

Before discussing the ML method of treating missing data, we will review some basic principles of ML estimates first.

Generally speaking, the basic principle of the ML method is to maximize the estimate of parameters, given statistical models and related data. More specifically, if we want to estimate a parameter θ, under the assumption that all observations are identically independently distributed (IID), the likelihood for the sample with *n* observations is

$$L(\theta) = \prod_{i=1}^{n} f(y_i|\theta),$$

where Π is a symbol for repeated multiplications.

From the equation above, θ is a set of unknown parameters that drive y_i. The problem then becomes to selects the values of the model parameters that maximize the probability of the observed data. In practice, it is common to work with the logarithm of L(θ), or ln(L(θ)), called the log-likelihood.

ML also can deal with incomplete-data in a similar way. For example, we plan to collect data on two variables, *x* and *y*, given a sample of *n* independent observations. For the first *m* observations, we can get data of both *x* and *y* variables; for the *n-m* observations, variable *x* is missing data and we only collect the data of variable *y*. For the observations with complete data, we present the likelihood by f (x, $y|\theta$), where θ is a set of unknown parameters that drive the distribution of *x* and *y*. Assume *x* is continuous, the likelihood for entire sample is like:

$$L(\theta) = \prod_{i=1}^{m} f(x_i, y_i|\theta) \prod_{i=n-m}^{n} f(y_i|\theta);$$

Of the many ways to obtain maximum likelihood estimators, the *Expectation-Maximization (EM) algorithm* is a very general method for estimating the parameters with missing data (Paul et al., 2001). It also can be used to obtain unbiased predictions for the missing values by parameters of the data model. As its name mentioned, *Expectation* indicates computing expected value of complete data, given observed data; *maximization* is maximizing the resulting function to obtain unknown parameters.

The process can be divided into 3 steps: first, one should estimate the unknown parameters, that is, the variance, covariance and means based on

complete data (*m* observations) and use them to build statistical models; then compute the missing values based on the statistical models. After all the missing values have been imputed, the new values of unknown parameters can be obtained by maximizing the probability of the entire dataset.

2.5.2 Bayesian Method of Treating Missing Data

An alternative approach of imputing missing data is called Bayesian method, which is widely used in dealing with categorical missing data. Bayesian computation in a missing data problem is based on the joint posterior distribution of parameters and missing data, given model assumptions and observed data.

In the Bayesian method, the estimate of distribution of the parameter θ can be made in terms of a probability statement, which is expressed as p(θ| y), and *y* is observed value. The joint probability distribution of θ and *y* can be expressed as a product of the prior distribution p (θ) and sampling distribution p(y|θ), respectively: p (θ, y) = p (θ) p(y|θ). Based on Bayes' rule, the posterior density is p (θ| y) $=\frac{p(\theta,y)}{p(y)} = \frac{p(\theta)p(y|\,\theta)}{p(y)}$. Since *y* indicates observed data, p(y) can be seen as a constant, yielding the posterior density as p (θ |y) ∝ p (θ) p (y|θ).

Chen and Astebro (2003) proposed an easy-to-implement Bayesian model-based approach to imputing missing categorical data, assuming the missing data mechanism is MCAR or MAR. If the missing data mechanism is MCAR, then we use the following procedure to impute missing data, under the assumption that the sampling model is a multinomial distribution with the parameter θ, and θ_{ij}=p(X=*i*,

Y=*j*|θ). If the missing data mechanism is MAR, we divide the value of X into *r* groups based on the values of X (X=1, 2… *r*) and classify observed Y into these groups. Then, we use the following procedure to impute missing data in each of these groups.

1. For a categorical variable Y with *k* possible outcomes, calculate the point estimate for the probability of each outcome *j* using the formula:

$\theta_j = \frac{y_j+1}{n+k}$, where y_j is the number of observations in the data for outcome *j*, and *n* is the total number of observations for variable Y.

2. Compute $P_j=\sum_{i=1}^{j} \theta_j$ for all *j* where and *j*=1 to *k*. Note that $P_k=1$.

3. For any observation *i* with a missing value, draw a random value, r_i from the range [0, 1). If $0 \leq r_i < P_1$, replace the missing value with outcome 1; if $P_1 \leq r_i < P_2$, replace the missing value with outcome 2; …; if $P_{k-1} \leq r_i$, replace the missing value with outcome *k*.

4. Repeat step 3 for all missing observations of the variable.

5. Repeat steps 1 through 4 for all other categorical variables with missing data satisfying MCAR to form a complete data set.

2.5.3 Multiple Imputations

The imputation methods that only one value is filled for each missing data point are called single imputation. Although they are strong techniques, single imputations cannot reflect the sampling variability in the actual values of the missing data under one model for non-response (Little & Rubin, 1987). Fortunately,

there is another alternative approach—multiple imputations (MI), which can remove these limitations.

Multiple imputations, was first proposed by Rubin in the early 1970's as a way to address missing data. Different from single imputation, where only one value is filled for each missing data point, multiple imputation indicates replacing each missing value by more than one imputed data.

Generally speaking, there are three main steps involved in multiple imputations (Paul et al., 2001). First, we repeat the imputation procedure more than once, producing multiple complete data sets. Second, each of these complete data sets is analyzed using standard analysis procedures. Third, the results from these complete data sets are combined for inference.

Rubin (1987) showed that the efficiency of an estimate based on *m* imputations, $(1 + \lambda/\mathrm{m})^{-1}$,where λ is the percentage of missing information. For instance, with 20% missing information, *m* = 4 imputations can achieve bout 95% efficiency.

III. Data Preprocessing

Two main phases are involved in this research: data preprocessing and model development. This chapter focuses on the first phase of the research, describing how the raw data retrieved from the ICGG Gaucher Registry are preprocessed before model development.

Data preprocessing is an important yet often neglected step in data mining, which includes data cleaning, data reduction, new data construction and data formatting. Data cleaning is necessary when dealing with incomplete, noisy and inconsistent data. The purpose of data reduction is to obtain a reduced dataset which is smaller yet contains most of the important information of the complete dataset. New data construction includes tasks such as generating new attributes and records, merging tables, and transforming data. Data formatting involves syntactic modifications to the data without changing its meaning. This step may be necessary for some particular modeling tools used in the next phase of model development.

3.1 Summary of Variables in the Raw Data

The data set from the ICGG Gaucher Registry includes 889 patients and 45 variables. Those variables can be divided into three categories based on their features. The first category includes 27 variables related to the patients' demographics and characteristics. The second category contains 8 variables related to the patients' hematological and visceral manifestations at their first infusion of imiglucerase. The remaining 10 variables are related to the patients' bone manifestations at their first infusion of imiglucerase. Tables 3 – 5 summarize meaning, type, value, and the number of missing values of each variable in the three categories, respectively.

Table 3. Variables Related to Each Patient's Demographics and Characteristics

Variable	Meaning	Type and values	Missing values
REG_ID_N	patient identification	categorical (unique ID for each patient)	0
AGEINF	patient's age at first infusion (in year)	continuous	0
INFUSDT	date of the first enzyme infusion	date-time	0
VISITDT	date of a DXA assessment visit	date-time	0
BISPHOS	whether the patient had treatment with bisphosphonates	binary (yes/no)	0
STARTDT	date of the treatment with bisphosphonates	date-time	0
YRSFUP	years between the first infusion and each DXA visit date	continuous	0
SEX	gender	binary(male/female)	0
ETHGRP	patient's ethnicity group	categorical (African-American; American-India; Asian; Caucasian; Hispanic; Jewish-Ashkenazi; Jewish-Both Ashkenzi and Sephardic; Jewish-Neither Ashkenzi or Sephardic; Jewish-Sephardic; Multi-Ethnic)	95 patients (10.7%)
REGION	patient's geographic region	categorical (Americas; Asia, Pacific, S.Africa; Europe; Middle East; USA)	0
GEN370S	patient's genotype	categorical (N370S/N370S;N370S/Other; Other/Other)	117 patients (13.2%)
SPLSTAT	patient's splenectomy status	binary (never/ever splenectomized)	1 patient (0.1%)
SPLDT	date of splenectomy	date-time	1 patient (0.1%)
DOSE3Y	average dose of imiglucerase (in U/kg/2wks)	continuous	24 patients (2.7%)
SPINEZ	lumbar spine DXA Z-score	continuous	0
DXASP2	whether a patient had more than one DXA assessment visit	binary(yes/no)	0
BSPINEZ	whether a patient had a record of baseline spine DXA score	binary(yes/no)	0
BSPINEZ1	whether a patient's baseline spine DXA Z-score is equal to or below -1	binary(yes/no)	0
APTOTYP	lumbar vertebrae total type	categorical (L1-L4;L2-L4)	5 patients (0.6%)
SPINETOT	lumbar vertebrae total L1 type	continuous	31 patients (3.5%)
DEXAMACH	measurement device	categorical (Hologic; Lunar)	98 patients (11.0%)
AGEGRP1	whether a patient's age is ≥5 to	binary (yes/no)	0

	<12		
AGEGRP2	whether a patient's age is ≥12 to <20	binary (yes/no)	0
AGEGRP3	whether a patient's age is ≥20 to <30	binary (yes/no)	0
AGEGRP4	whether a patient's age is ≥30 to <50	binary (yes/no)	0
BMIB	a patient's body mass index at baseline	continuous	292 patients (32.8%)

Table 4. Variables Related to Patients' Hematological and Visceral Manifestations

Variable	Meaning	Type and Value	Missing Values
HGB	hemoglobin(G/DL)	continuous	204 patients (22.9%)
ANEM	whether a person has anemia	binary (yes/no)	204 patients (22.7%)
PLT	platelet count(×10^3/mm^3)	continuous	204 patients (22.7%)
THROM	thrombocytopenia category	categorical (none/mild; moderate; severe)	202 patients (22.7%)
SPLENO	splenomegaly category	categorical (none/mild; moderate; severe)	496 patients (55.8%)
SPLMN	spleen volume(multiples of normal)	integer	496 patients (55.8%)
HEPATO	hepatomegaly category	categorical (none/mild; moderate; severe)	493 patients (47.6%)
LIVMIN	liver volume (multiple of normal)	integer	493 patients (47.6%)

Table 5. Variables Related to Patients' Bone Manifestations

Variable	Meaning	Type and Value	Missing Values
INFARC	presence of infraction	binary(yes/no)	603 patients (67.8%)
EFD	presence of erlenmeyer flask deformity	binary(yes/no)	565 patients (63.6%)
AVN	presence of avascular necrosis	binary(yes/no)	572 patients (64.3%)
MARR	presence of marrow infiltration	binary(yes/no)	525 patients (59.1%)
FRACT	presence of fractures	binary(yes/no)	702 patients (79%)
LYTIC	presence of lytic lesions	binary(yes/no)	702 patients (79%)

OSTEO	presence of osteopenia	binary(yes/no)	633 patients (71.2%)
BPAINPM	presence of bone pain during past month	binary(yes/no)	251 patients (28.2%)
BPAINSEV	levels of bone pain severity	categorical(very mild, mild, moderate, severe, extreme)	711 patients (80.0%)
BCRISLS	presence of bone crises since last submission	binary(yes/no)	471 patients (46.9%)

3.2 Imputation of Missing Data

As shown in Tables 3 – 5, most variables in the raw data set have missing values. Therefore, we need to decide how to handle the missing data before model development. In this study, we imputed the missing values for three variables of patients' demographics and characteristics which are considered important potential predictive variables for model development. They are ETHGRP, DOS3Y, and BMIB.

In particular, we used the maximum likelihood (ML) method, which we mentioned in chapter 2, to impute missing values of DOSE3Y based on REGION, SEX, AGEINF, and SPINEZ at the patient's first DXA assessment visit, and impute the missing values of BMIB based on ETHGRP, SEX, and AGEINF.

Although the ML based imputation method is powerful, it also has limitations. Because the ML method assumes that the variables involved are normally distributed, it cannot handle categorical variables effectively. Therefore, we used an alternative approach—Bayesian imputation to impute the missing values of ETHGRP based on REGION and GEN370S.

Because ETHGRP and BMIB have relatively high missing rates – 10.7% and 32.8%, respectively, we carried out multiple imputations for them. Based on the formula of multiple imputations described in chapter 2, $(1 + \lambda/m)^{-1}$, where λ is the percentage of missing information and *m* is the number of imputations, we imputed ETHGRP twice and BMIB six times in order to achieve about 95% efficiency in their estimates.

3.3 New Variable Construction

Although imputation is a good method of handling missing data, it cannot handle variables with too many missing values, such as the variables related to the patient's hematological, visceral, and bone manifestations shown in Table1. In order to solve this problem, we constructed 11 new variables based on the original variables, as summarized in Table 6.

Table 6. Summary of Constructed New Variables

New Variables	Meaning	Type
SPINEZ_FIRST	the patient's DXA Z-score at his/her first DXA assessment visit	numeric
BIWEEK_FIRST_VISIT	the number of bi-weeks between the patient's first infusion and first DXA visit date	numeric
BIWEEK_FOLLOW_VISIT	the number of bi-weeks between the patient's first DXA assessment visit and each of his/her follow-up DXA assessment visits	numeric
BIWEEK_CONVISIT	the number of bi-weeks between the patient's consecutive visits	numeric
RATE_SPINEZ	the rate of change of the patient's DXA Z-score from his/her immediately previous visit to the current visit, calculated as the patient's DXA Z-score in the current visit subtracted from that in his/her immediately previous visit, and then divided by the number of bi-weeks between the two consecutive visits (only if the patient had multiple visits)	numeric

FIRST_RATE_SPINEZ	the rate of change of the patient's DXA Z-score from his/her first infusion to current visit	numeric
PREVIOUS_SPINEZ	the DXA Z-score in each patient's immediately previous DXA assessment visit	numeric
THROM_HEPATO	Combine THROM and HEPATO to see how many of them were categorized as "severe".	non-negative Integer(0, 1, 2)
SPLSTAT_SPLENO	Combine SPLSTAT and SPLENO to see whether SPLSTAT is "ever splenectomized" or SPLENO is "severe". If so, SPLSTAT_SPLENO is assigned as "severe", otherwise, it is "not severe"	binary (severe/not severe)
BONE_PROBLEMS	Combine INFARC, EFD, AVN, MARR, FRACT, LYTIC, and OSTEO to see how many of these variables were reported as "yes" for a patient.	non-negative Integer (0,1,2,3, 4,5,6,7)
BONE_PAIN	Combine BPAINPM, BPAINSEV, and BCRISLS to see whether a patient's BPAINPM was recorded as "yes" and BPAINSEV as "severe" or "extreme", or BCRISLS was recorded as "yes"; if so, BONE_PAIN was assigned "yes", otherwise, "no".	binary(yes/no)

3.4 Distributions of Selected Variables

In this section, we present the distributions of 13 variables. These variables are AGEINF, BISPHOS, SEX, ETHGRP, REGION, DOSE3Y, SPINEZ_FIRST, BMIB, BIWEEK_FIRST_VISIT, BIWEEK_FOLLOW_VISIT, BIWEEK_CONVISIT, FIRST_RATE_SPINEZ, and RATE_SPINEZ. For ETHGRP, DOSE3Y and BMIB, because we have imputed their missing values, their distributions before and after the imputations are both illustrated in this section.

Figure 1 illustrates the distribution of AGEINF. From this figure, we can see the largest age group in the data set was between 5 and 10 years old. Figure 2 illustrates the distribution of BISPHOS, which shows most patients (750 out of 891 patients) did not have treatment with bisphosphonates. Figure 3 illustrates the

distribution of SEX, which shows that there were more female (594 out of 891 patients) than male patients. Figure 4 illustrates the distribution of REGION, which shows that most patients were from the USA (452 of 891 patients), followed by Europe (214 of 891 patients), the Middle East (107 of 891 patients), the Americas (102 of 891 patients) and Asia, Pacific and South Africa (16 of 891 patients).

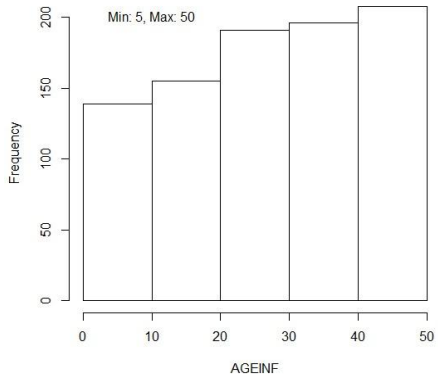


Figure 1. Histogram of AGEINF

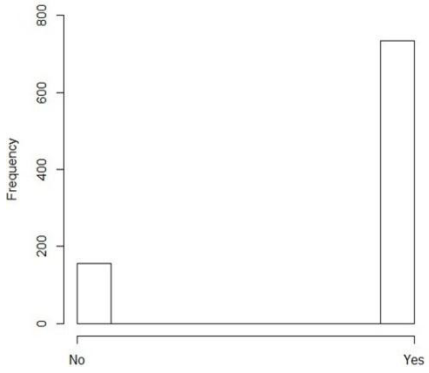


Figure 2. Histogram of BISPHOS

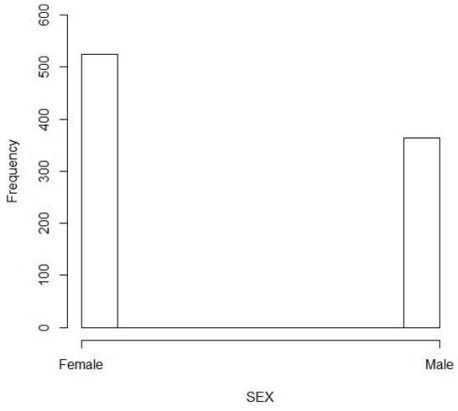


Figure 3. Histogram of SEX

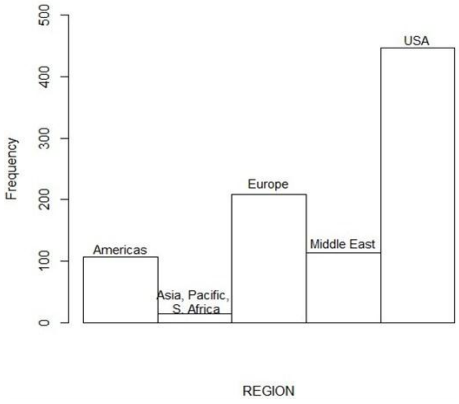


Figure 4. Histogram of REGION

Figure 5(a) shows the distribution of ETHGRP with missing data. We can see that the two largest ethnic groups were Caucasian and Jewish Ashkenazi, which together accounted for about 80% of all the patients, followed by Hispanic and African American, which accounted for about 5% and 4% of the total number of patients, respectively. Each of the remaining groups had less than 1% of all the patients.

Figures 5(b) and (c) illustrate the distributions of ETHGRP after the 1st and 2nd imputations, respectively. We can see their distributions are very similar to the distribution before imputation (Figure 4); they also show that more than 80% of the patients were Ashkenazi Jewish and Caucasian.

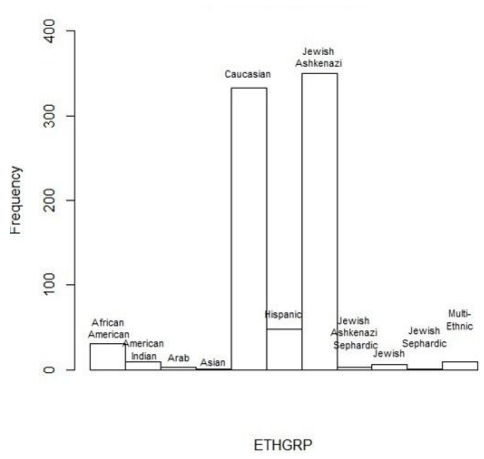


Figure 5 (a). Histogram of ETHGRP before Imputation

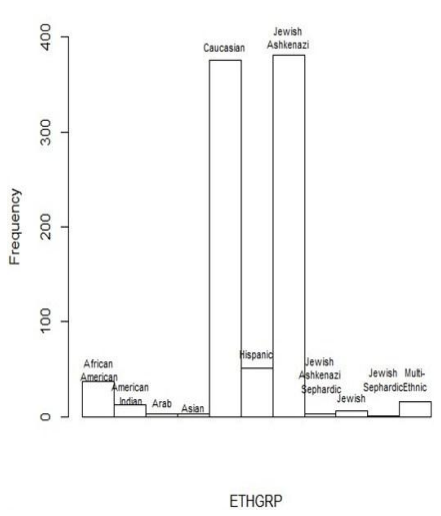


Figure 5 (b). Histogram of ETHGRP after the 1st Imputation

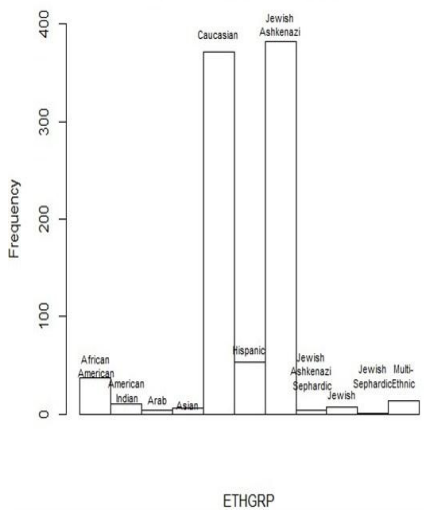


Figure 5 (c). Histogram of ETHGRP after the 2nd Imputation

Figures 6(a) and (b) illustrate the distributions of DOSE3Y before and after imputation, respectively. They both indicate that most patients' dosages were between 50 and 60 U/kg/2wks, between 20 and 30 U/kg/2wks, and between 10 and 20 U/kg/2wks.

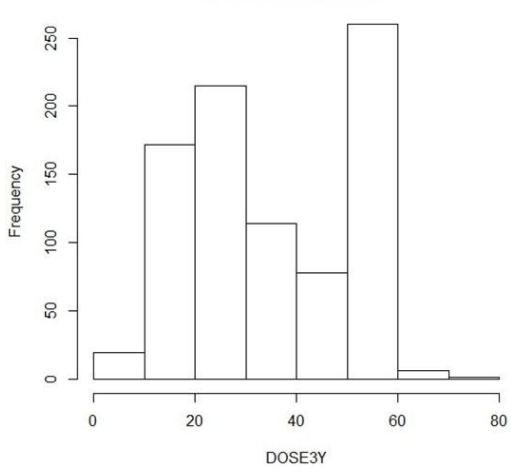


Figure 6(a). Histogram of DOSE3Y before Imputation

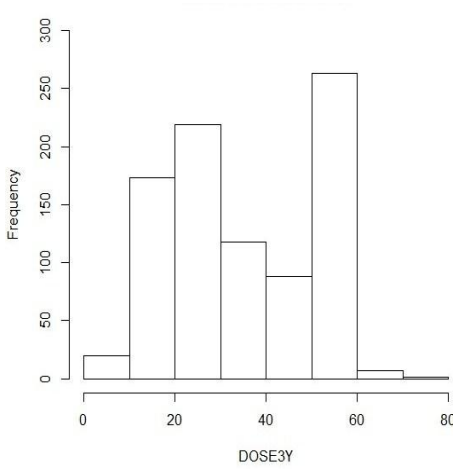


Figure 6(b). Histogram of DOSE3Y after Imputation

Figure 7(a) illustrates the distribution of BMIB with missing data. We can see that the BMIB values of most patients ranged between 10 and 30. Figures 7(b) and 7(c) illustrate the distributions of BMIB after the 1st and 2nd imputations, respectively. In total, there are 12 imputations (2 imputations of ETHGRP $\times 6$ imputations of BMIB) of missing data. Since all of their distributions are very similar, only two of them are illustrated here. Comparing Figures 7(a), (b), and (c), we can see that some of the imputed values of BMIB are below 10 whereas the original BMIB values were always greater than 10. However, all 3 figures show that most patients' BMIB values fell between 10 and 30.

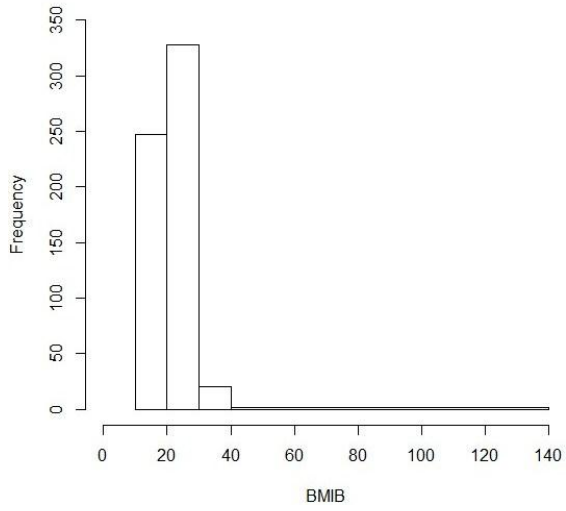


Figure 7(a). Histogram of BMIB before Imputation

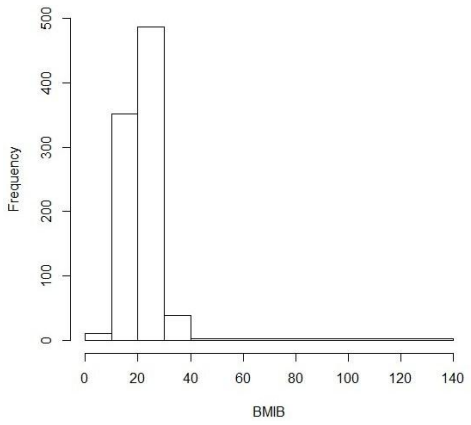


Figure 7(b). Histogram of BMIB after 1st Imputation

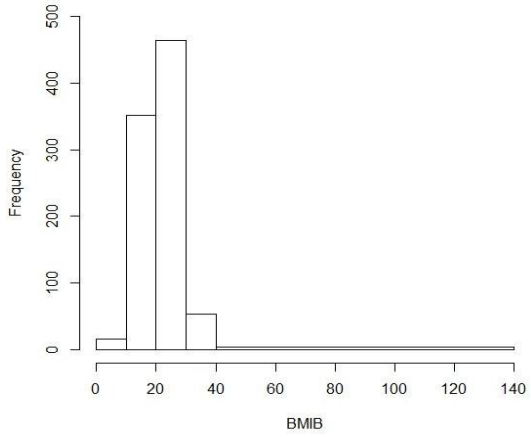


Figure 7(c). Histogram of BMIB after 2nd Imputation

Figure 8 illustrates the distribution of BIWEEK_FIRST_VISIT, which suggests that the first DXA assessment visits of most patients occurred within 250 bi-weeks after their initial infusions. Negative values of BIWEEK_FIRST_VISIT mean that the first DXA assessment visits of some patients occurred before their initial infusions.

Figure 8 illustrates the distribution of BIWEEK_FOLLOW_VISIT, which shows us that the most follow-up DXA assessment visits occurred within 250 bi-weeks of the first assessment visit, especially during the first 150 bi-weeks.

Figure 9 illustrates the distribution of BIWEEK_COVISIT, from which we can see most time intervals between a patient's two consecutive visits were less than 50 bi-weeks, followed by between 50 and 100 bi-weeks, and between 100 and 150 bi-weeks. Very few time intervals were greater than 150 bi-weeks.

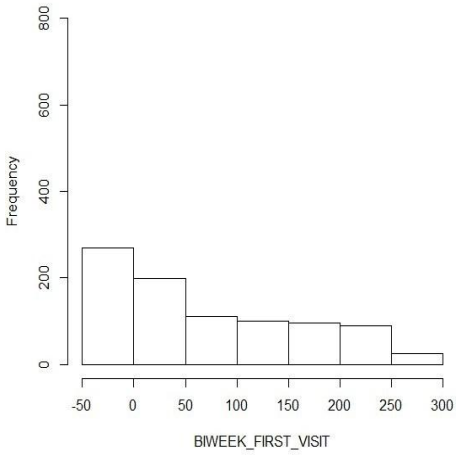


Figure 8. Histogram of BIWEEK_FIRST_VISIT

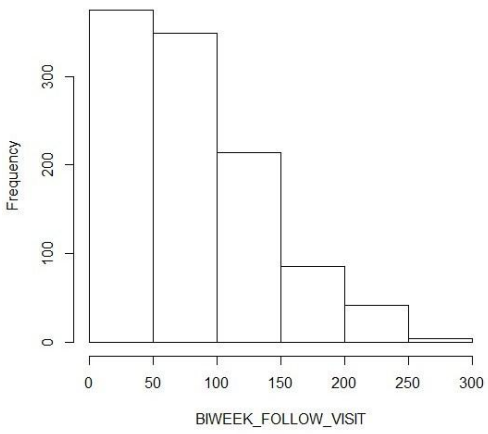


Figure 9. Histogram of BIWEEK_FOLLOW_VISIT

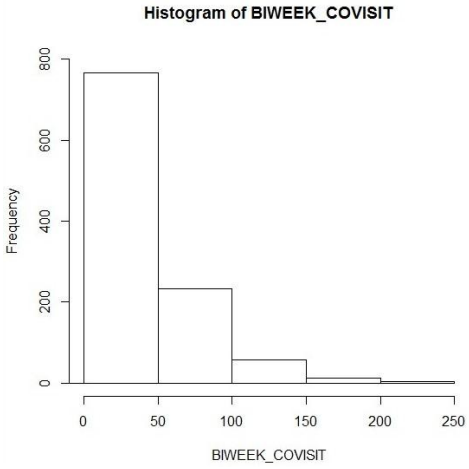


Figure 10. Histogram of BIWEEK_COVISIT

Figure 11 illustrates the distribution of SPINEZ_FIRST, which approximates a normal distribution, with most values ranging between -2 and 0.

Figure 12 illustrates the distribution of FIRST_RATE_SPINEZ, which suggests that most rates of change of a patient's DXA Z-score from that of the first DXA assessment visit were between -0.1 and 0.1.

Figure 13 illustrates the distribution of RATE_SPINEZ, which demonstrates a similar pattern as that in Figure 12, with most values ranging from -0.1 to 0.1,

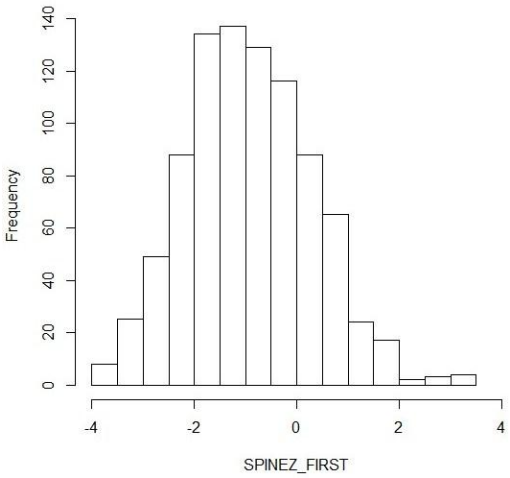


Figure 11. Histogram of SPINEZ_FIRST

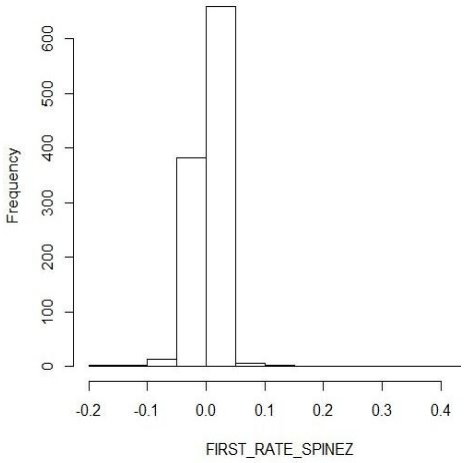


Figure 12. Histogram of FIRST_RATE_SPINEZ

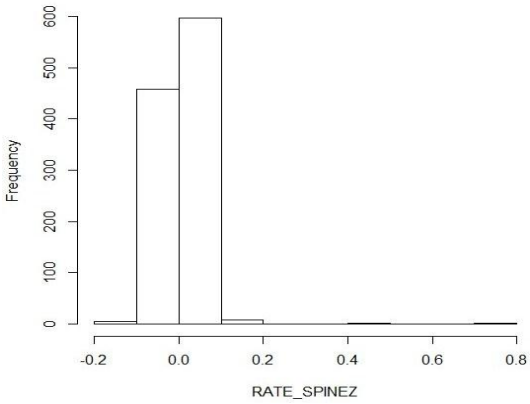


Figure 13. Histogram of RATE_SPINEZ

IV. Model Development

In this project, we focus on building models for three response variables by using the regression tree method: SPINEZ_FIRST, which represents a patient's DXA Z-score at his/her first DXA assessment visit, FIRST_RATE_SPINEZ, which is the rate of change of the patient's DXA Z-score from his/her first infusion to the current DXA assessment visit, and RATE_SPINEZ, which refers to the rate of change of the patient's DXA Z-score between two consecutive DXA assessment visits. For each response variable, models have been derived from the 12 complete datasets with imputed missing values, as described in chapter 3 of the thesis, as well as from the original data set with missing values.

4.1 Regression Tree Method

Regression tree methods are used in this research to derive models from the preprocessed data thanks to the two main advantages of regression trees over traditional regression methods. First, it is very easy to interpret the results of a regression tree model, because the final results of a regression model can be summarized in a series of logical "if-then" conditions. Second, regression trees are nonparametric and nonlinear, making no implicit assumption about the underlying relationship between the response variable and covariates.

In particular, two regression tree algorithms with different approaches to selecting the splitting covariate and stopping criteria are applied in this project. The

first algorithm is the RPART package in R (Atkinson et al., 2000), which implements many of the ideas of CART. The other algorithm is the PARTY package in R, which selects the splitting criteria based on the theory of permutation test and conditional distribution of statistics.

4.1.1 RPART Package

In RPART, the splitting attribute, x_j^R, is selected based on minimization of expected sum of variances for two resulting nodes, as shown in Equation 1:

$$\arg\min \quad [p_l \text{var}(Y_l) + p_r \text{var}(Y_r)] \qquad (1)$$

$$x_j \leqslant x_j^R,\ j=1,\ldots, M$$

where p_l and $p_r$represent the probabilities of cases in the left and right nodes, respectively. var (Y_l) and var (Y_r) represent the variances of response variable Y in the left and right child nodes, respectively. This splitting criterion is equivalent to maximizing the between-group sum-of-squares and minimizing the within-group sum-of-squares in the analysis of variance (ANOVA).

Like CART, RPART uses the cost-complexity function to select the best tree size. Given a tree T with |T| leaf nodes T_1, $T_2 \ldots T_{|T|}$ and the cost of adding one more leaf node into the model α, the cost of T is defined as follows:

$$R\alpha(T) = R(T) + \alpha|T| \qquad (2)$$

where

$$R(T) = \text{risk of } T = \sum_{i=1}^{|T|} P(T_i)R(T_i) \qquad (3)$$

in which $P(T_i)$ and $R(T_i)$ represent the probability and risk of leaf node T_i, respectively. By these definitions, selecting a tree with the "optimal" size is finding the smallest tree T for which $R\alpha$ (T) is minimized. Cross-validation is used to choose the best value for α.

4.1.2 PARTY Package

Like most regression tree algorithms, RPART has a selection bias towards covariates with many possible splits. The "ctree" routine in the PARTY package addresses this problem by separating the selection of splitting covariates and splitting points of the covariates into two distinct steps at each iteration of the tree construction. Measuring the association between covariates and the response variable by conditional distribution of statistics is the basis for unbiased selection among covariates measured at different scales (Hothorn et al, 2006).

Given the responsible variable Y and *m* covariates ($X_1, \ldots, X_m$), the conditional distribution Y given X is:

$$D(Y|X) = D(Y|X_1, \ldots, X_m) \quad (4)$$

To test whether there is dependency between Y and X_j ($j = 1, 2 \ldots m$), the null hypothesis is

$$H_0^j: D(Y|X_j) = D(Y), \quad (5)$$

The global null hypothesis is

$$H_0 = \cap_{j=1}^{m} H_0^j, \quad (6)$$

If H_0 cannot be rejected at a pre-specified significance level α, the recursion of the tree is stopped. Otherwise, the covariate with the strongest association to Y

is selected as the splitting covariate. The association between Y and X_j(j=1… *m*) is measured by a form of linear statistics:

$$T_j(L_n, w)=\text{vec}(\sum_{i=1}^{n} w_i g_j(X_{ji})h(Y_i, (Y_1, \dots, Y_n))^T) \qquad (7)$$

where w_i is the case weight of *i*, g_j is a non-random transformation of the covariate X_j, *h* is the influence function, a non-random transformation of Y. Selection of g_j and *h* depends on the distributions of X_j and Y. The distribution of $T_j(L_j, w)$ depends on the joint distribution of Y and X_j.

To find the distribution of $T_j(L_j, w)$, permutation tests on each possible permutation of the response, S (L_j, w) are used. The conditional expectation μ_j and covariance $\sum j$ are:

$$\mu_j=E\ (T_j(L_j, w)|\ S\ (L_j, w)) \qquad (8)$$

$$\sum j=V\ (T_j(L_j, w)|\ S\ (L_j, w)) \qquad (9)$$

Then, we calculate the p-value of the conditional test for H_0^j on each covariate. If the minimum of the p-values is less than a pre-specified nominal level α, H_0 is rejected, and the covariate with the minimum p-value is selected as a splitting attribute. Test of the global hypothesis H_0 can be based on univariate p-values or multiple test procedures such as Bonferroni-adjusted p-values and the min-p-value resampling approach.

After the splitting covariate x_{j^*} is selected, the next step is to find the best split of x_{j^*} by evaluating all possible splits. The two-sample linear statistic, which

measures the discrepancy between the two disjoint subsets created by a split of X_j^*, A and X_j^*/A, is calculated as follows:

$$T_{j^*}^{A}(L_j, w)=\mathrm{vec}\left(\sum_{i=1}^{n} w_i\, I\left(X_{j^*i}\epsilon A\right)h(Y_i, (Y_1, \ldots, Y_n))^{T}\right) \qquad (10)$$

where I (.) is the indicator function. The best split is the one that maximizes the test statistic over all possible subsets A.

4.2 Modeling Results of SPINEZ_FIRST

The first response variable of the study is SPINEZ_FIRST, the patient's DXA Z-score at his/her first DXA assessment visit. We used both the RPART and PARTY packages to build models of SPINEZ_FIRST, using 10-fold cross-validation in RPART models and the Bonferroni-adjusted p-value of 0.95 in PARTY models. The minimum number of records in each leaf node is set to 20 (about 2% of all the records in the data set). The 12 variables used to build the models of SPINEZ_FIRST are summarized in Table 7.

Table 7. Variables Used to Build the Models of SPINEZ_FIRST

Variable	Meaning	Type and Value
AGEINF	patients' age at first infusion(in years)	numeric
BISPHOS	whether the patient had treatment with bisphosphonates	binary (yes/no)
SEX	gender	binary (male/female)
ETHGRP	patient's ethnicity group	categorical (African-American; American-Indian; Arab; Asian; Caucasian; Hispanic; Jewish- Ashkenazi; Jewish- Both Ashkenazi and Sephardic; Jewish-

		Neither Ashkenazi nor Sephardic; Jewish- Sephardic; Multi-Ethnic)
REGION	patient's geographic region	categorical (Americas; Asia, Pacific, S. Africa; Europe; Middle East; USA)
BIWEEK_FIRST_VISIT	the number of bi-weeks between first infusion and first DXA assessment visit	numeric
DOSE3Y	average dose of imiglucerase (in U/kg/2wks)	numeric
BMIB	a patient's body mass index at the first infusion	numeric
THROM_HEPATO	combine THROM and HEPATO to see how many of them were categorized as "severe"	numeric (0, 1, 2)
SPLSTAT_SPLENO	combine SPLSTAT and SPLENO to see whether SPLSTAT is “ever splenectomized” or SPELNO is “severe”	binary (severe/not severe)
BONE_PROBLEMS	combine INFARC, EFD, AVN, MARR, FRACT, LYTIC and OSTEO to see how many of these variables were reported “yes” for a patient	numeric (0,1,2,3,4,5,6)
BONE_PAIN	combine BPAINPM, BPAINSEV, and BCRISLS to see whether for a patient, his/her BPAINPM was recorded as“yes” and BPAINSEV as "severe" or “extreme”, or his/her BCRISLS was recorded as “yes”. if so, BONE_PAIN was assigned “yes”, otherwise, it is “no”	binary(yes/no)

According to the modeling results of SPINEZ_FIRST, the most predictive variables are REGION, BISPHOS, SEX, and BIWEEK_FIRST_VISIT although other variables included in different models vary. All the models of SPINEZ_FIRST derived using the PARTY package are the same, as they do not include any variable with missing values. The models of SPINEZ_FIRST derived from the RPART package, on the other hand, are similar for the complete datasets with the same imputed ETHGRP values but different for those with different

imputed BMIB values. We present 3 models derived using the RPART package from 3 complete datasets, as well as the model derived from the original dataset using the RPART package. The other 9 tree models derived in the study are reported in the appendix.

4.2.1 Models of SPINEZ_FIRST by Using PARTY package

Figure 14 illustrates the regression tree model of SPINEZ_FIRST developed using the PARTY package, in which the 3 numbers, from top to bottom, underneath each leaf node show the number of records, the mean, and standard deviation of SPINEZ_FIRST of the node, respectively. The model suggests that for the patients who had bisphosphonates, female patients had significantly higher values of SPINEZ_FIRST than male patients (t-test, $p=0.0002$). Among the patients who didn't have bisphosphonates, those in Asia, Pacific, S.Africa, Europe and the USA had significantly higher values of SPINEZ_FRIST than those in other regions (t-test, $p<0.0001$). Additionally, patients who had large values of BIWEEK_FIRST_VISIT (the number of bi-weeks between the patients' first infusion visits and their first DXA assessment visits) had significantly higher values of SPINEZ_FIRST than those with smaller values of BIWEEK_FIRST_VISIT (t-test, $p=0.0006$).

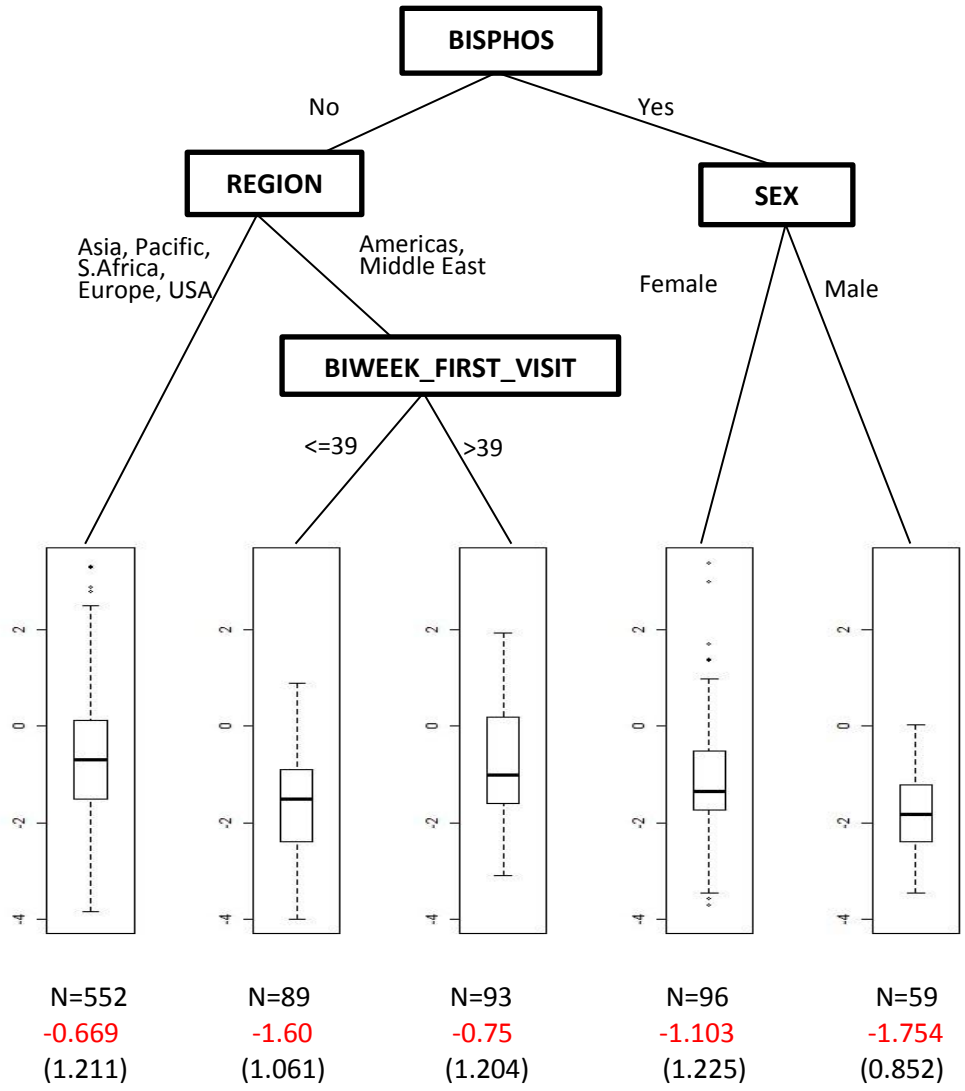


Figure 14. Model of SPINEZ_FIRST Derived Using the PARTY Package from All the Individual Complete Datasets and the Original Dataset

(The 3 numbers, from top to bottom, underneath each leaf node, show the number of records, the mean, and standard deviation of SPINEZ_FIRST of the node, respectively.)

4.2.2 Models of SPINEZ_FIRST Using the RPART Package

Figure 15 illustrates the first regression tree model of SPINEZ_FIRST developed using the RPART package. Compared to Figure 14, Figure 15 includes

more variables in the tree model. First, patients were divided into two groups based on REGION. Among the patients in Europe, those who had higher values of BIWEEK_FIRST_VISIT have significantly higher SPINEZ_FIRST values than others. Among the patients who resided in the other regions and had bisphosphonates, female patients had significantly higher values of SPINEZ_FIRST than male patients. For the patients who did not have bisphosphonates and whose values of BIWEEK_FIRST_VISIT were below 159.5, their values of SPINEZ_FIRST values were affected by ETHGRP, AGEINF and BMIB. More specifically, African American, Asian, both Ashkenazi and Sephardic Jewish, neither Ashkenazi nor Sephardic Jewish, and multi-ethnic patients had significantly smaller SPINEZ_FIRST values than those in other ethnicity groups (t-test, p=0.0003). Additionally, among the American Indian, Arab, Hispanic, Caucasian, Jewish-Ashkenazi, and Jewish-Sephardic patients, those younger than 11 years old had significantly higher values of SPINEZ_FIRST than older patients (t-test, p=0.0006), and those whose body mass index at their first infusion were higher than 22.82 had significantly higher value of SPINEZ_FIRST than others (t-test, p=0.0004).

Figure 16 illustrates the second regression tree model of SPINEZ_FIRST developed using the RPART package from a complete dataset with the same imputed values of ETHGRP but with different imputed BMIB values, as those in Figure 15. The major difference between Figures 15 and 16 is that BMIB has more detailed partitions given AGEINF. Like Figure 15, Figure 16 also suggests that the patients with higher value of BMIB had larger value of SPINEZ_FIRST.

Figure 17 illustrates the third regression tree model of SPINEZ_FIRST developed using the RPART package from a complete dataset with different imputed ETHGRP values and different imputed BMIB values as those in Figure 15. There are two major differences between Figures 15 and 17. First, for the patients whose BIWEEK_FIRST_VISIT values were no greater than 101.5, their SPINEZ_FIRST values were affected by their values of BMIB, particularly those with larger values of BMIB had significantly higher values of SPINEZ_FIRST. Second, the SPINEZ_FIRST values of female patients were affected by REGION. More specifically, the patients in Europe had significantly lower SPINEZ_FIRST values than the patients in other regions (t-test, p=0.0003).

Figure 18 illustrates the tree model of SPINEZ_FIRST developed using the RPART package from the original dataset. It is very similar to Figure 17, except for one main difference: neither Ashkenazi nor Sephardic Jewish and Multi-Ethnic patients are grouped together with American Indian, Arab, Caucasian, Hispanic, Sephardic Jewish, and Ashkenazi Jewish patients in Figure 18, rather than with African American, Asian, and both Ashkenazi and Sephardic Jewish patients, as shown in Figure 17.

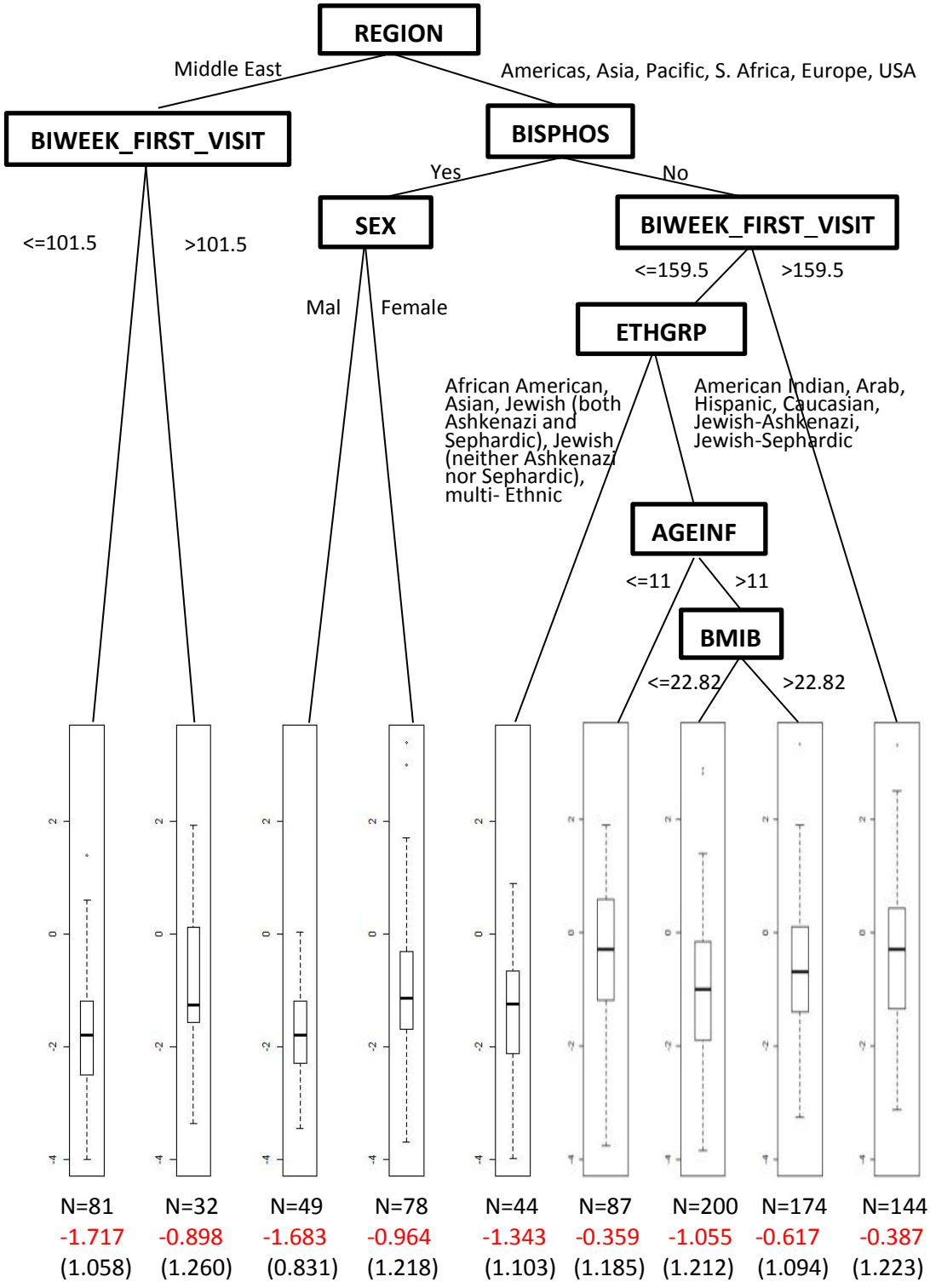


Figure 15. Model 1 of SPINEZ_FIRST Derived Using the RPART Package from a Complete Dataset

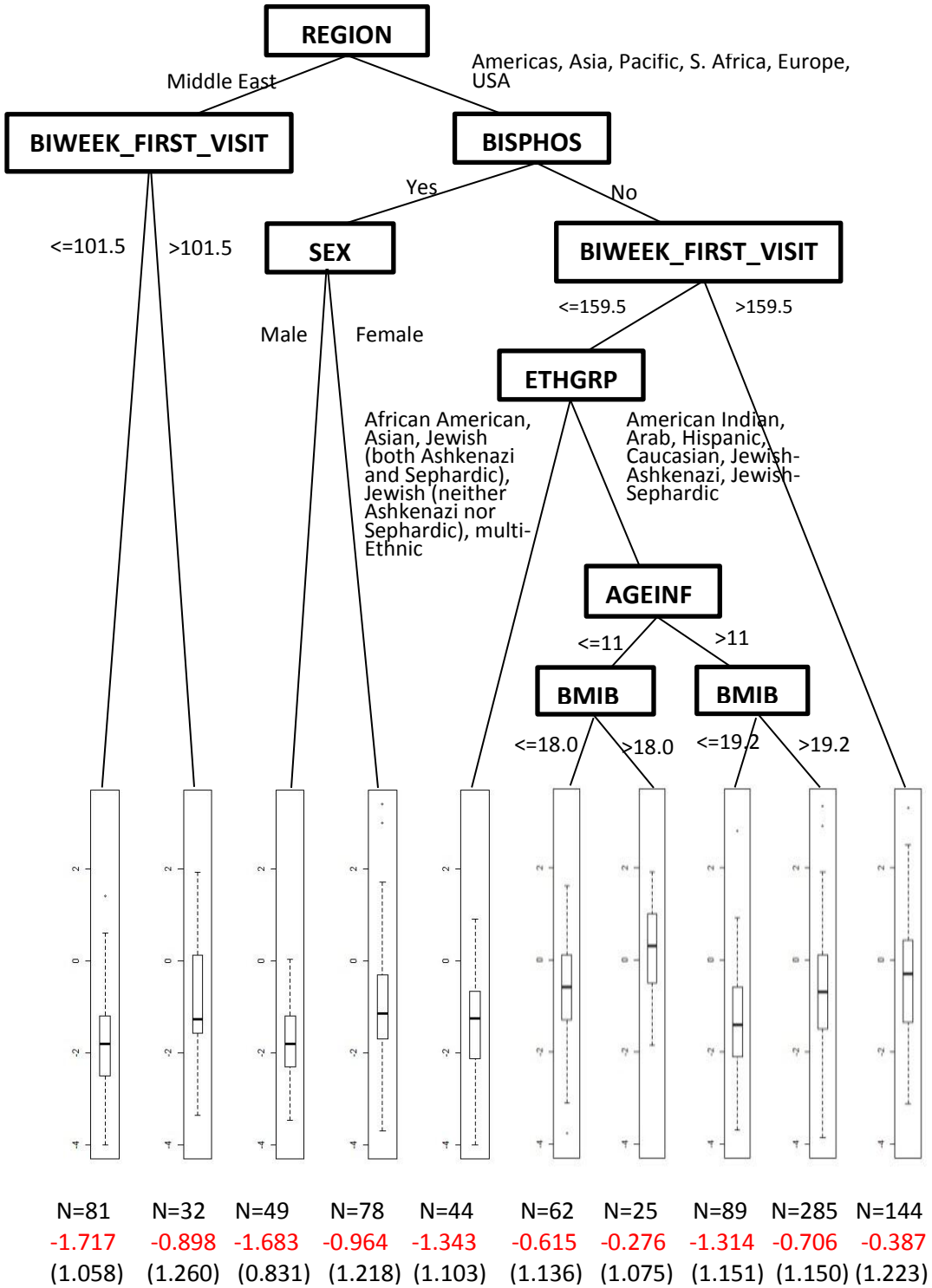


Figure 16. Model 2 of SPINEZ_FIRST Derived Using the RPART Package from a Complete Dataset

(The complete dataset has the same imputed ETHGRP values as those used in Figure 15 but with different imputed BMIB values.)

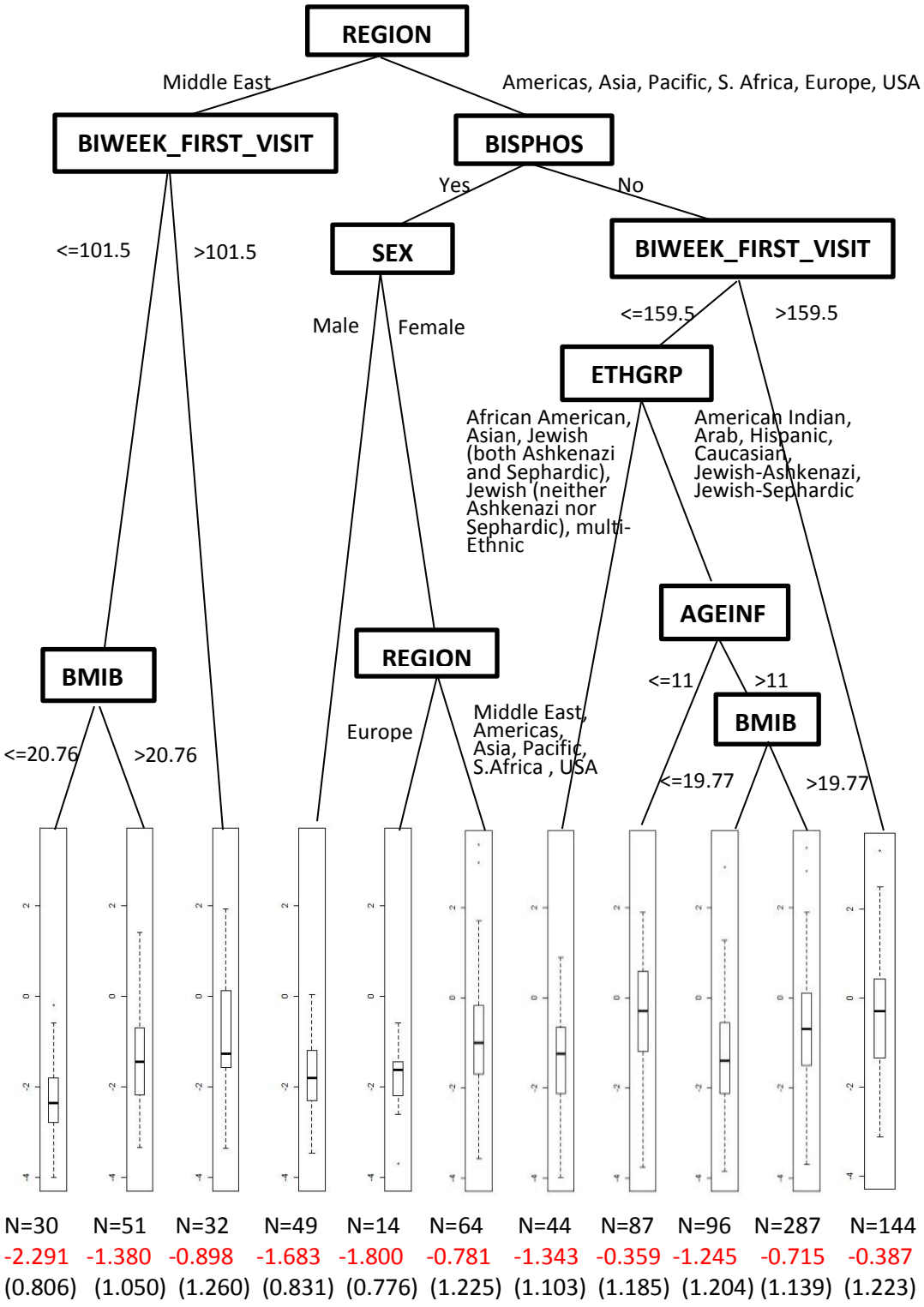


Figure 17. Model 3 of SPINEZ_FIRST Derived Using the RPART Package from a Complete Dataset

(The complete dataset has different imputed ETHGRP and BMIB values from those used in Figures 15 and 16.)

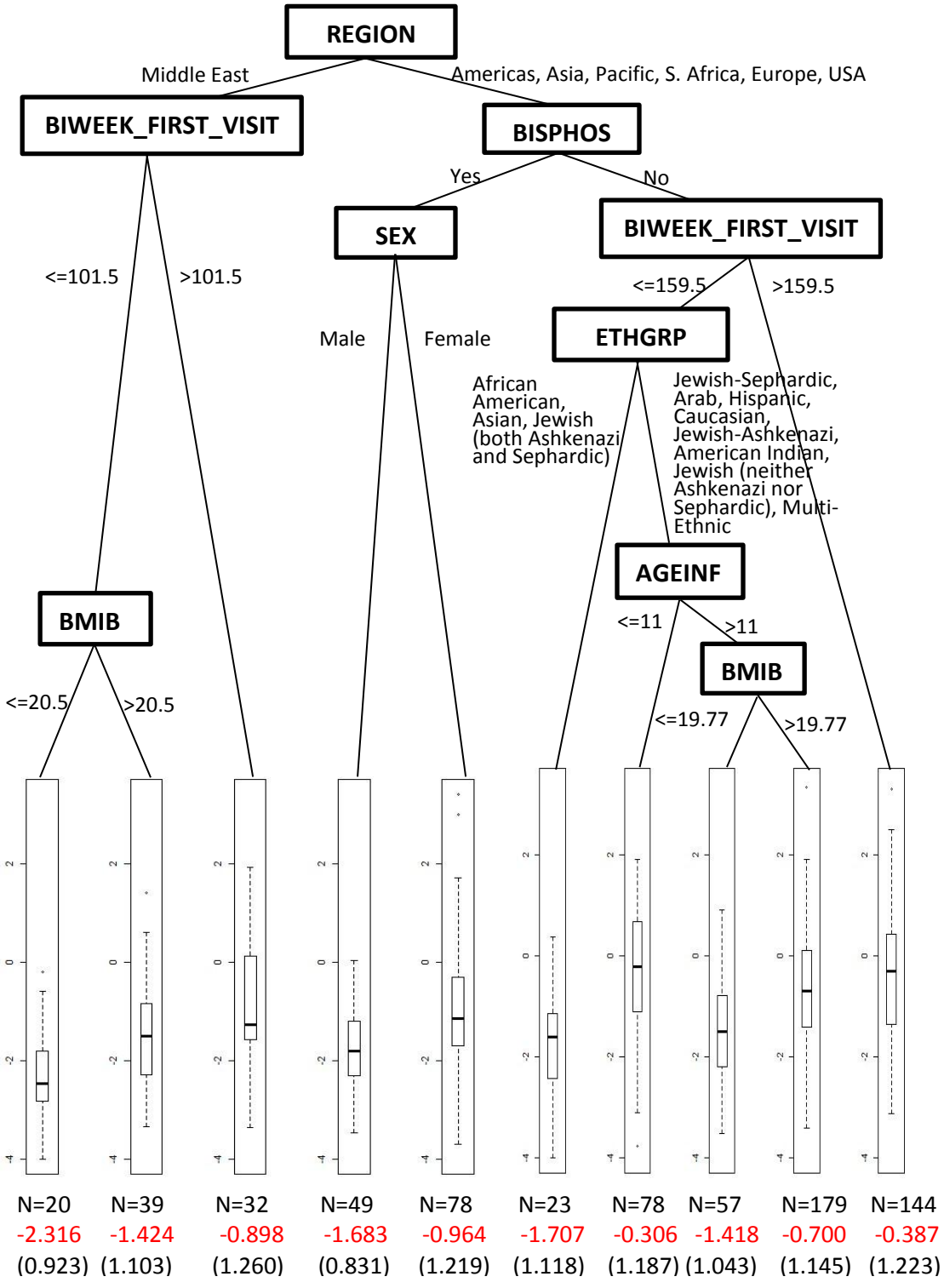


Figure 18. Model 4 of SPINEZ_FIRST Derived Using the RPART Package from the Original Dataset

4.3 Modeling Results of FIRST_RATE_SPINEZ

Only the patients who had at least two DXA assessment visits are used to build models of FIRST_RATE_SPINEZ. In addition, 3 outliers whose values of FIRST_RATE_SPINEZ were greater than 0.15 or less than -0.15 are removed. As a result, only data from 471 patients with 1066 records are used to build models of FIRST_RATE_SPINEZ.

The variables used to build the models of FIRST_RATE_SPINEZ are summarized in Table 8. In addition to the variables used to build the models of SPINEZ_FIRST, as listed in Table 7, the response variable discussed in section 4.2 is also included as a predictor in building the models of FIRST_RATE_SPINEZ. Besides, FIRST_BIWEEK_UP, the number of bi-weeks between a patient's first DXA assessment visit and each of his/her follow-up DXA assessment visits, is another new variable in the models of FIRST_RATE_SPINEZ.

Table 8. The Variables Used in Building Models of FIRST_RATE_SPINEZ

Variable	Meaning	Type and Value
AGEINF	patients' age at first infusion(in years)	numeric
BISPHOS	whether the patient had treatment with bisphosphonates	binary (yes/no)
SEX	gender	binary (male/female)
ETHGRP	patient's ethnicity group	categorical (African-American; American-Indian; Arab; Asian; Caucasian; Hispanic; Jewish- Ashkenazi; Jewish- Both Ashkenazi and Sephardic; Jewish-Neither Ashkenazi nor Sephardic; Jewish- Sephardic; Multi-Ethnic)
REGION	patient's geographic region	categorical (Americas;

		Asia, Pacific, S. Africa; Europe; Middle East; USA)
BIWEEK_ FOLLOW_VISIT	the number of bi-weeks between each patient's first DXA Z-Score visit and each of the following-up DXA Z-Score visits	numeric
DOSE3Y	average dose of imiglucerase (in U/kg/2wks)	numeric
BMIB	a patient's body mass index at the first infusion	numeric
THROM_HEPATO	combine THROM and HEPATO to see how many of them were categorized as "severe"	numeric (0, 1, 2)
SPLSTAT_SPLENO	combine SPLSTAT and SPLENO to see whether SPLSTAT is "ever splenectomized" or SPELNO is "severe"	binary (severe/not severe)
BONE_PROBLEMS	combine INFARC, EFD, AVN, MARR, FRACT, LYTIC, and OSTEO to see how many of these variables were reported "yes" for a patient	numeric (0,1,2,3,4,5,6)
BONE_PAIN	combine BPAINPM, BPAINSEV, and BCRISLS to see whether for a patient, his/her BPAINPM was recorded as"yes" and BPAINSEV as "severe" or "extreme", or his/her BCRISLS was recorded as "yes". if so, BONE_PAIN was assigned "yes", otherwise, it is "no"	binary(yes/no)
SPINEZ_FIRST	the DXA Z-Score at each patient's first DXA assessment visit.	numeric

4.3.1 Models of FIRST_RATE_SPINEZ Using PARTY Package

Figure 19 illustrates the model of FIRST_RATE_SPINEZ developed using the PARTY package from the 12 complete datasets and from the original dataset, which suggests that the patients with higher SPINEZ_FIRST values had significantly lower FIRST_RATE_SPINEZ values than those with lower SPINEZ_FIRST values. In particular, most patient whose SPINEZ_FIRST values were greater than 0.36 had negative FIRST_RATE_SPINEZ values, and the patients in Europe had significantly lower FIRST_RATE_SPINEZ values than

those in other regions (t-test, $p<0.0001$). For the patients in Americas, Asia, Middle East, USA, Pacific, and S. Africa, whose SPINEZ_FIRST values ranged from -2.7 to 0.36, their values of FIRST_RATE_SPINEZ were related to SEX, AGEINF, BIWEEK_ FIRST_VISIT, BISPHOS, and ETHGRP. In particular, Asian, Caucasian, Ashkenazi Jewish, both Ashkenazi and Sephardic Jewish, neither Ashkenazi nor Sephardic Jewish, and Multi-Ethnic patients had significantly smaller FIRST_RATE_SPINEZ values than the patients in other ethnic groups (t-test, $p<0.0001$). Among these patients, those who had bisphosphonates had significantly lower FIRST_RATE_SPINEZ values than those without (t-test, $p<0.0001$). Male patients had significantly lower values of FIRST_RATE_SPINEZ than female patients (t-test, $p<0.0001$). Among these female patients, those with lower values of BIWEEK_ FOLLOW_VISIT tended to have higher values of FIRST_RATE_SPINEZ.

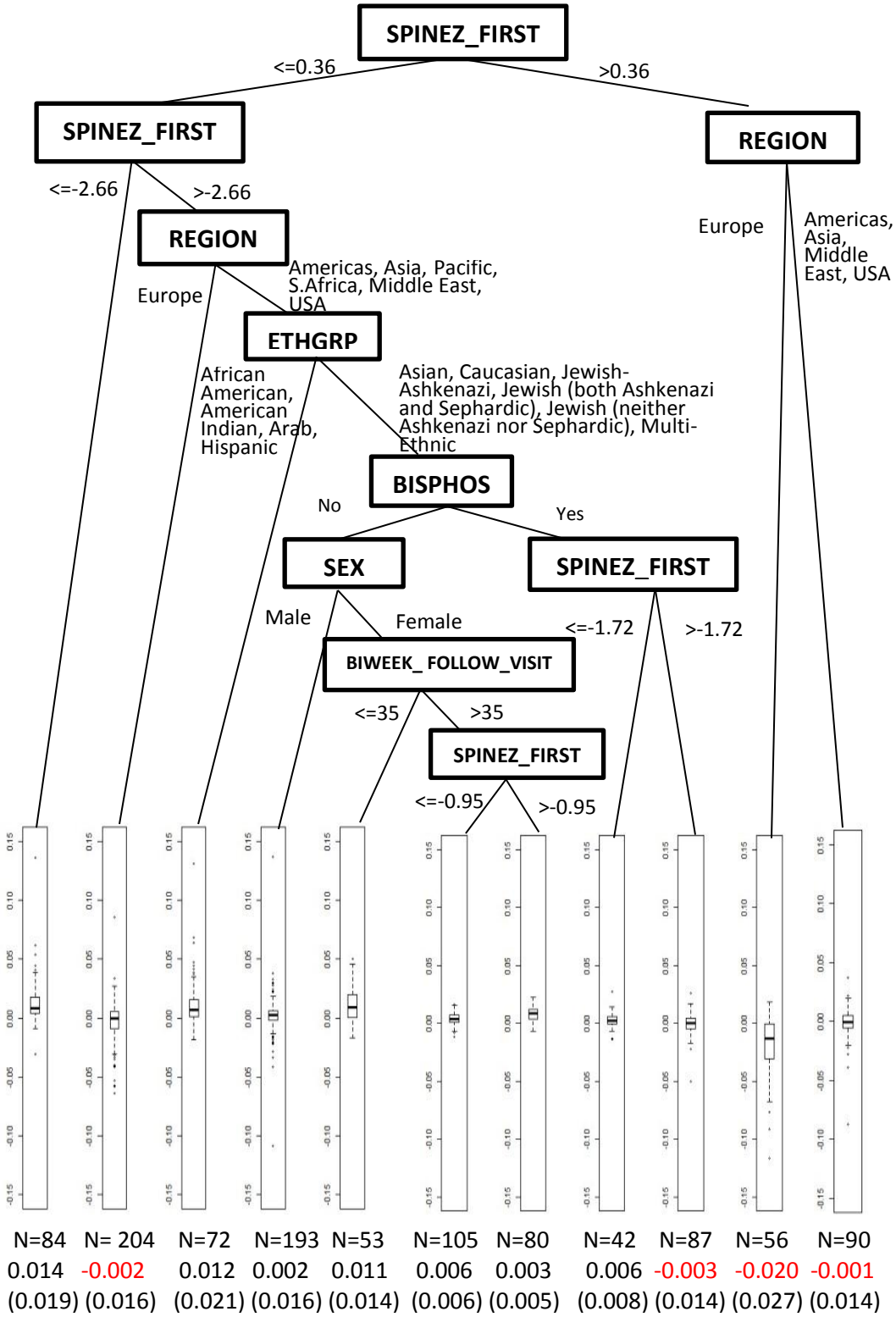


Figure 19. Model of FIRST_RATE_SPINEZ Derived Using the PARTY Package from the 12 Complete Datasets and from the Original Dataset.

4.3.2 Models of FIRST_RATE_SPINEZ Using RPART Package

Figure 20 illustrates the model of FIRST_RATE_SPINEZ developed using the RPART package from the 12 complete datasets. From the figure, we can see that SPINEZ_FIRST and REGION are two most predictive variables of FIRST_RATE_SPINEZ. Most patients whose SPINEZ_FIRST values were larger than 0.36 had negative FIRST_RATE_SPINEZ values. For the patients whose SPINEZ_FIRST values ranged from -2.7 to 0.36, their values of FIRST_RATE_SPINEZ were related to ETHGRP and REGION. More specifically, African American, American Indian, Arab, and Hispanic patients had significantly higher FIRST_RATE_SPINEZ values than the other patients (t-test, $p<0.0001$). Among the Asian, Caucasian, Ashkenazi Jewish, both Ashkenazi and Sephardic Jewish, neither Ashkenazi nor Sephardic Jewish, and multi-Ethnic patients, those in Europe had significantly lower values of FIRST_RATE_SPINEZ than those in other regions (t-test, $p<0.0001$).

Figure 21 illustrates the model of FIRST_RATE_SPINEZ developed using the RPART package from the original dataset. It is very similar to Figure 20, except for one difference: compared to Figure 20, Figure 21 had an additional predictor BMIB in the model. Among the patients in the Americas, Asia, Pacific, S. Africa, Middle East and the USA, those whose BMIB values were greater than 15.5 had significantly higher value of FIRST_RATE_SPINEZ than those in other regions (t-test, $p<0.0001$).

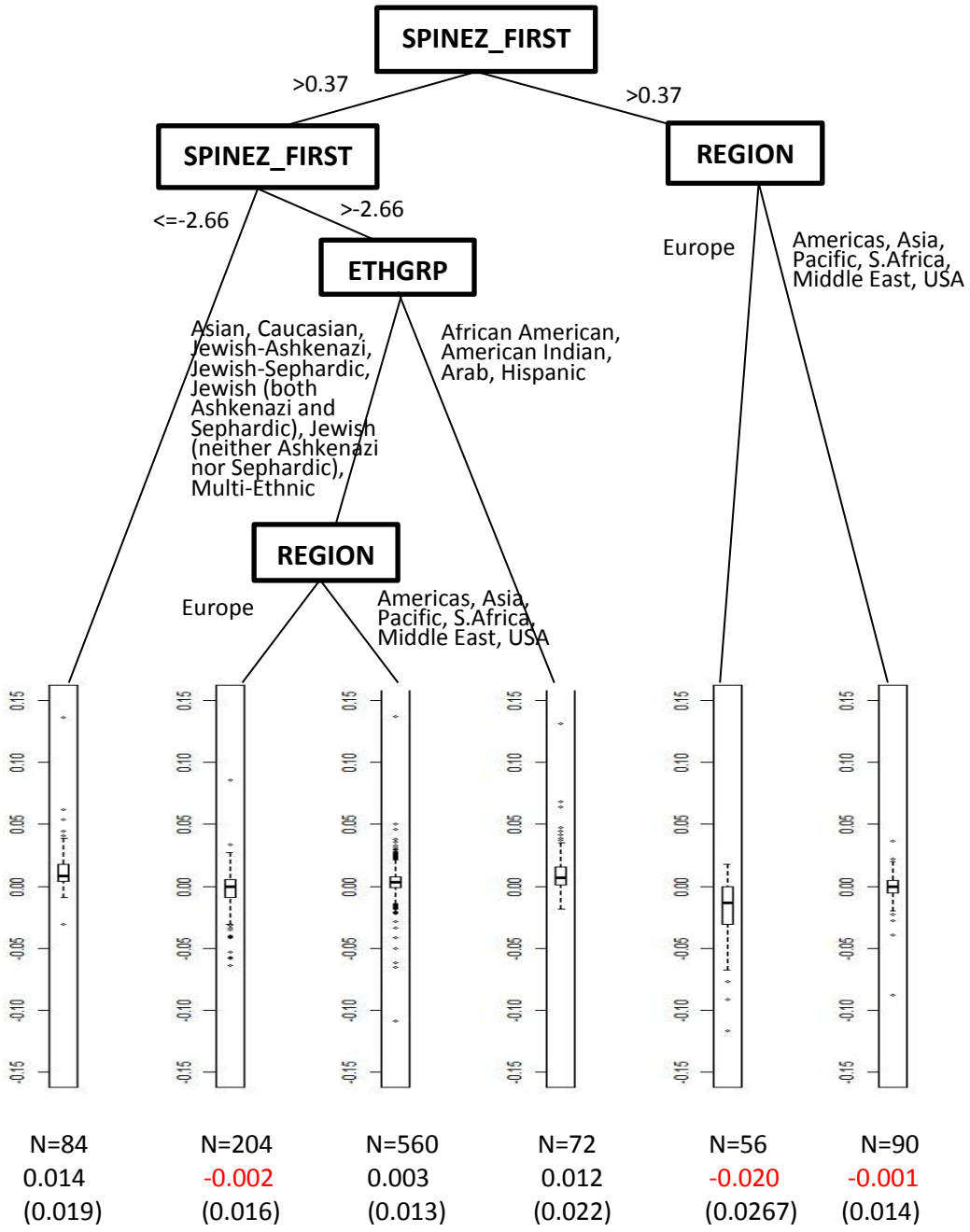


Figure 20. Model of FIRST_RATE_SPINEZ Derived Using the RPART Package from All the Complete Datasets

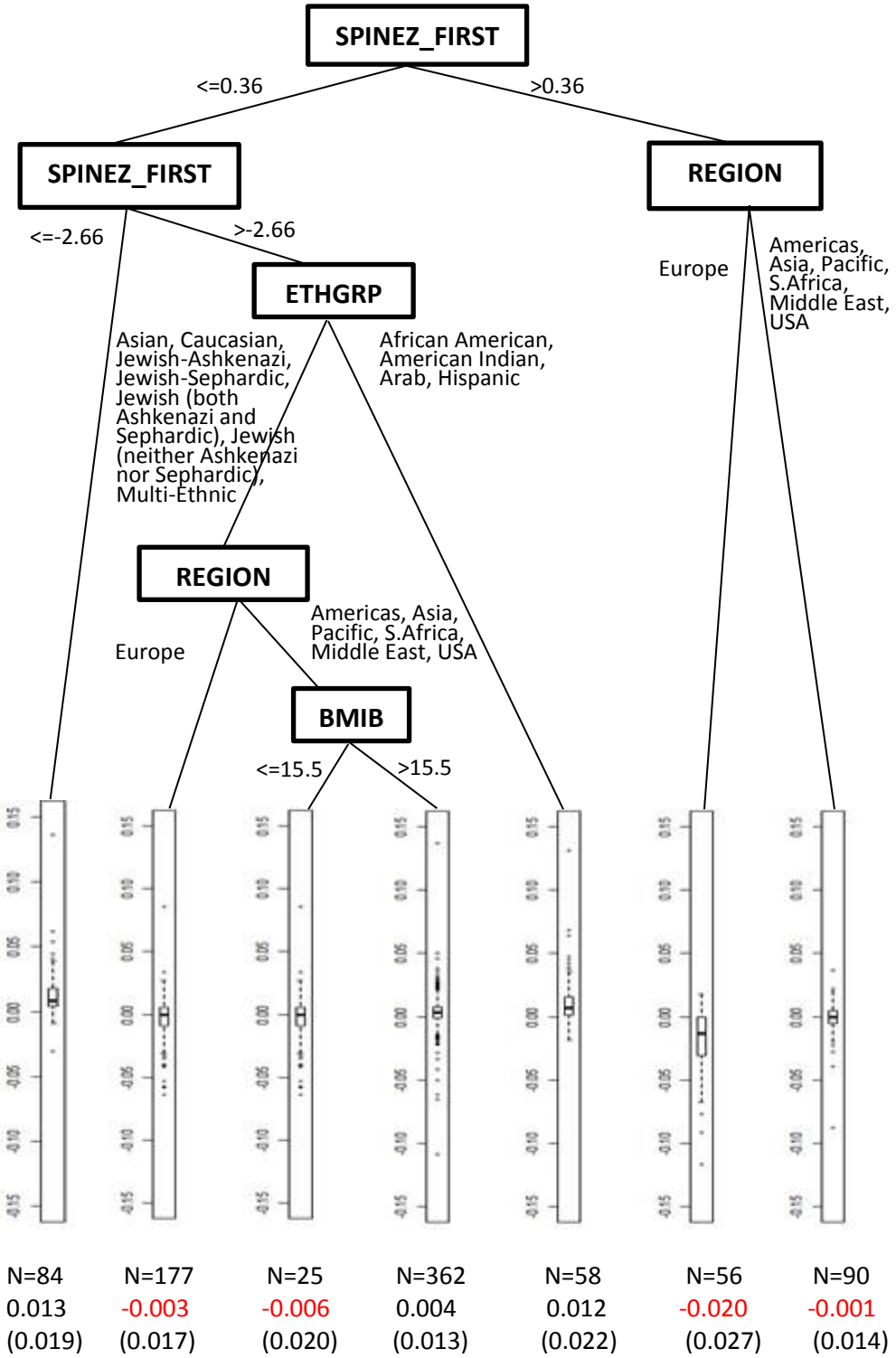


Figure 21. Model of FIRST_RATE _ SPINEZ Derived Using the RPART Package from the Original Dataset

4.4 Modeling Results of RATE_SPINEZ

Like the dataset used to build the models of FIRST_RATE_SPINEZ, the dataset used to build the models of RATE_SPINEZ also only contains the patients who had multiple DXA assessment visits. Besides, 2 outliers whose RATE_SPINEZ values are larger than 0.2 or smaller than -0.2 are removed. As a result, data from 471 patients with 1067 records are used to derive models of RATE_SPINEZ.

Table 9 summarizes the variables used to build the models of RATE_SPINEZ. Compared to Table 8, Table 9 does not include FIRST_BIWEEK_UP or SPINEZ_FIRST, but it contains PREVIOUS_ SPINEZ, which represents the patient's DXA Z-score at his/her immediately previous DXA assessment visit, and BIWEEK_COVISIT, which refers to the number of bi-weeks between the patient's two consecutive DXA assessment visits.

Table 9. Variables Used to Build Models of RATE_SPINEZ

Variable	Meaning	Type and Value
AGEINF	patients' age at first infusion (in years)	numeric
BISPHOS	whether the patient had treatment with bisphosphonates	binary (yes/no)
SEX	gender	binary (male/female)
ETHGRP	patient's ethnicity group	categorical (African-American; American-Indian; Arab; Asian; Caucasian; Hispanic; Jewish- Ashkenazi; Jewish- Both Ashkenazi and Sephardic; Jewish-Neither Ashkenazi nor Sephardic; Jewish- Sephardic; Multi-Ethnic)
REGION	patient's geographic region	categorical (Americas;

		Asia, Pacific, S. Africa; Europe; Middle East; USA)
BIWEEK_COVISIT	the number of bi-weeks between two successive DXA Z-Score visits	numeric
DOSE3Y	average dose of imiglucerase (in U/kg/2wks)	numeric
BMIB	a patient's body mass index at the first infusion	numeric
THROM_HEPATO	combine THROM and HEPATO to see how many of them were categorized as "severe"	numeric (0, 1, 2)
SPLSTAT_SPLENO	combine SPLSTAT and SPLENO to see whether SPLSTAT is “ever splenectomized” or SPELNO is “severe”	binary (severe/not severe)
BONE_PROBLEMS	combine INFARC, EFD, AVN, MARR, FRACT, LYTIC and OSTEO to see how many of these variables were reported “yes” for a patient	numeric (0,1,2,3,4,5,6)
BONE_PAIN	combine BPAINPM, BPAINSEV, and BCRISLS to see whether for a patient, his/her BPAINPM was recorded as“yes” and BPAINSEV as "severe" or “extreme”, or his/her BCRISLS was recorded as “yes”. if so, BONE_PAIN was assigned “yes”, otherwise, it is “no”	binary(yes/no)
PREVIOUS_SPINEZ	the DXA Z-Score in each patient’s immediately previous DXA assessment visit	numeric

The models of RATE_SPINEZ developed using the PARTY and RPART packages are much simpler than the models of the other two response variables. All the models of RATE_SPINEZ suggest that PREVIOUS_SPINEZ is the most predictive variable of RATE_SPINEZ.

4.4.1 Models of RATE_SPINEZ Developed Using the PARTY Package

Figure 22 illustrates the model of RATE_SPINEZ developed using the PARTY package from all of the 12 complete datasets and from the original dataset, which suggests that PREVIOUS_SPINEZ is the only important predictive variable

of RATE_SPINEZ. Particularly, the lower the PREVIOUS_SPINEZ, the higher the RATE_SPINEZ.

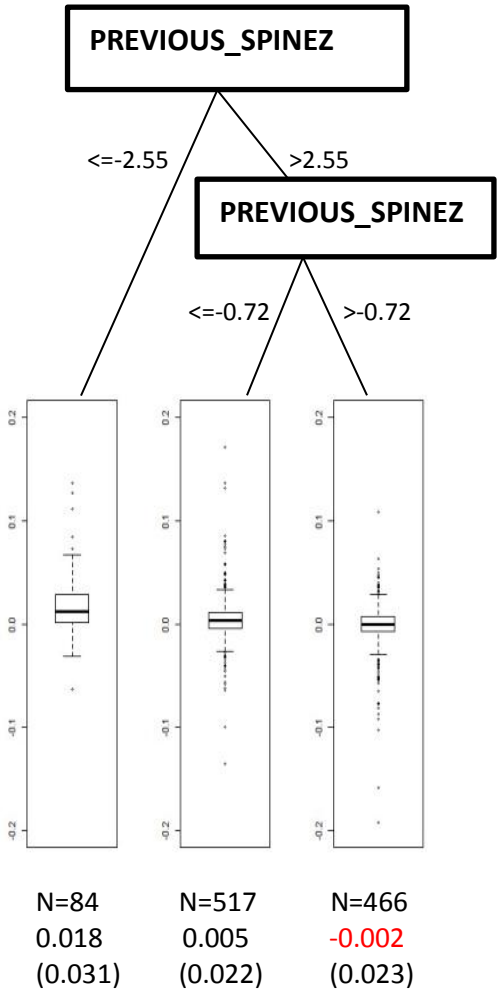


Figure 22. Model of RATE_SPINEZ Derived Using the PARTY Package from All the Individual Complete Datasets and from the Original Dataset

4.4.2 Models of RATE_SPINEZ Developed Using the RPART Package

Figure 23 illustrates the model of RATE_SPINEZ developed using the RPART package from all the complete datasets. It suggests that most patients

whose BIWEEK_COVISIT values were less than 14.5 had negative values of RATE_SPINEZ. Otherwise, for the patients whose BIWEEK_COVISIT values were greater than 14.5, and PREVIOUS_SPINEZ values were less than -2.595, their values of RATE_SPINEZ are affected by ETHGRP. More specifically, African American, Ashkenazi Jewish, and Multi-Ethnic patients had significantly lower values of RATE_SPINEZ than the other patients (t-test, $p<0.0001$). For the patients in the other ethnicity groups, whose BIWEEK_COVISIT values were more than 14.5, their RATE_SPINEZ values were only related to PREVIOUS_SPINEZ: the lower the PREVIOUS_SPINEZ, the higher the RATE_SPINEZ.

Figure 24 illustrates the model developed from the original dataset using the RPART package. It is almost identical to Figure 23, except with one difference: ETHGRP in Figure 23 is replaced by BMIB in Figure 24. In particular, the patients with BMIB values less than 22.5 had significantly lower values of RATE_SPINEZ than those with higher BMIB values.

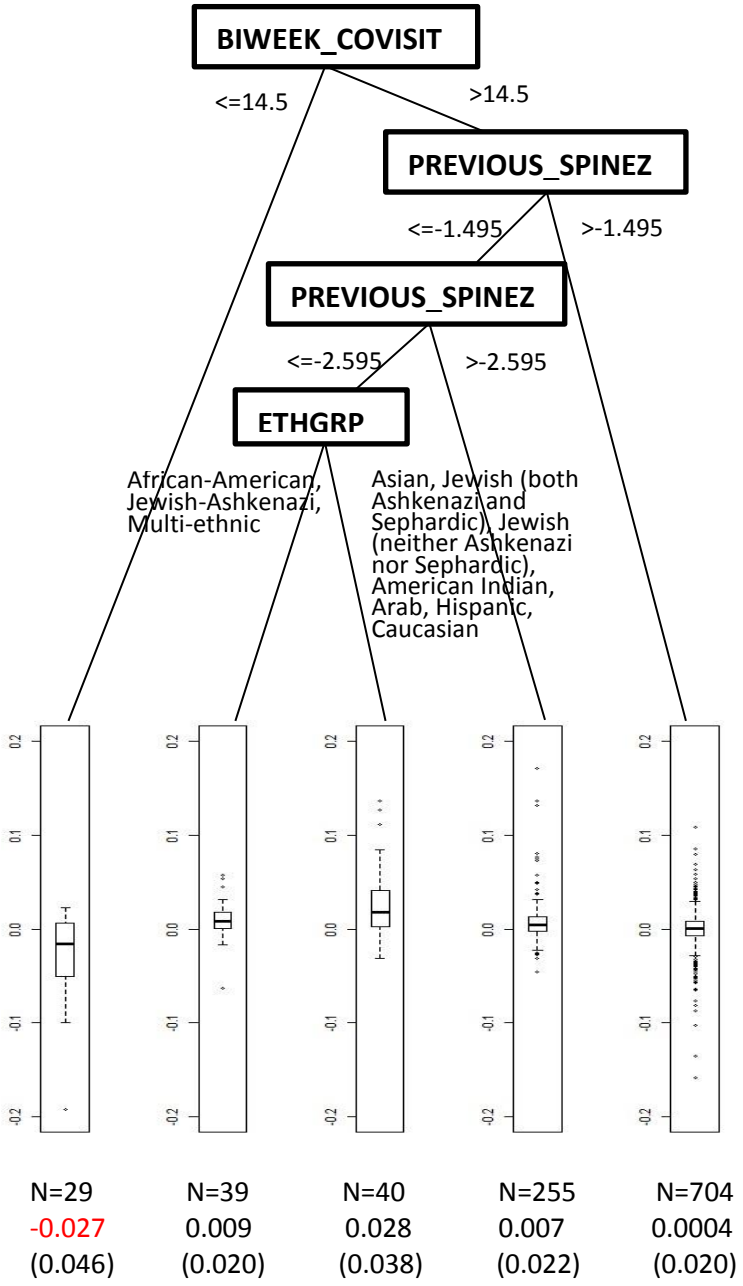


Figure 23. Model of RATE_SPINEZ Derived Using the RAPRT Package from All the Complete Datasets

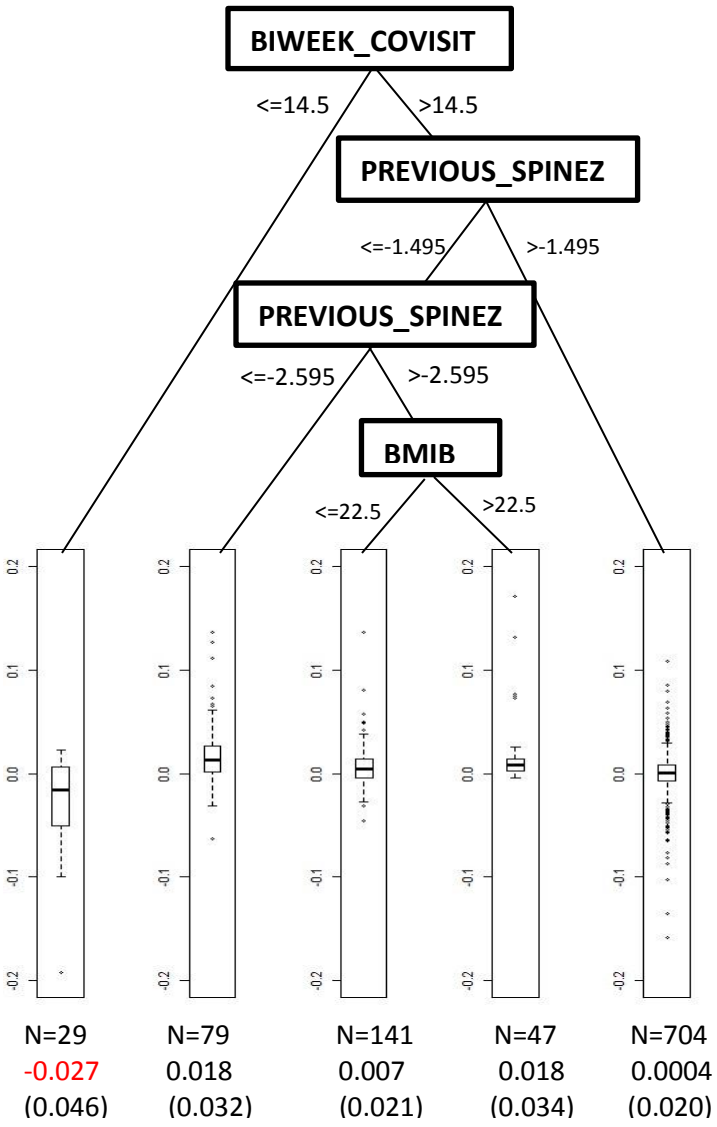


Figure 24. Model of RATE_SPINEZ Derived Using the RAPRT Package from the Original Dataset

4.5 Summary of Key Findings

We summarize the key findings from the derived models of SPINEZ_FIRST, FIRST_RATE_SPINEZ and RATE_SPINEZ.

4.5.1 Findings of SPINEZ_FIRST

All the models of SPINEZ show that the most predictive variables of SPINEZ_FIRST are REGION (the patient's geographic region), BISPHOS (whether the patient had bisphosphonates), SEX (patients' gender) and BIWEEK_UP_FIRST (the number of bi-weeks between the patient's first infusion and his/her first DXA assessment visit). More details are as follows:

- Patients who had larger values of BIWEEK_UP_FIRST had significantly larger values of SPINEZ_FIRST than those with smaller values of BIWEEK_UP_FIRST.
- Among the patients in Middle East, their values of SPINEZ_FIRST were affected by BIWEEK_UP_FIRST. Those with larger values of BIWEEK_UP_FIRST had higher values of SPINEZ_FIRST.
- Among the patients who had bisphosphonates, their values of SPINEZ_FIRST were affected by SEX. Particularly, female patients had significantly higher values of SPINEZ_FIRST than male patients.

4.5.2 Models of FIRST_RATE_SPINEZ

All the models of FIRST_RATE_SPINEZ suggest that the most predictive variables of FIRST_RATE_SPINEZ are REGION, SPINEZ_FIRST, and ETHGRP. More details are listed as follows:

- Patients with smaller SPINEZ_FIRST values had significantly higher values of FIRST_RATE_SPINEZ than the patients with larger SPINEZ_FIRST values.
- Patients with SPINEZ_FIRST less than -2.66 had the largest FIRST_RATE_SPINEZ, while most patients with SPINEZ_FIRST greater than 0.36 had negative FIRST_RATE_SPINEZ.
- For the patients with SPINEZ_FIRST ranging between -2.66 and 0.36, their values of FIRST_RATE_SPINEZ were affected by REGION and ETHGRP. Those in Europe had significantly lower values of FIRST_RATE_SPINEZ than those in other regions. In addition, African American, American Indian, Arab, and Hispanic patients had significantly larger values of FIRST_RATE_SPINEZ than the other patients.

4.5.3 Models of RATE_SPINEZ

All the models of RATE_SPINEZ suggest that the most predictive variables of RATE_SPINEZ is PREVIOUS_SPINEZ. In particular, patients with smaller PREVIOUS_SPINEZ had greater RATE_SPINEZ than those with larger PREVIOUS_SPINEZ.

V. Discussion and Conclusions

In this research, we have applied two regression tree methods – RPART and PARTY packages in R – to build models for three response variables:

SPINEZ_FIRST, FIRST_RATE_SPINEZ, and RATE_SPINEZ. For each response variable, models have been derived from 12 complete datasets that contain variations of imputed values for missing covariates, as well as from the original data set with missing values. The regression trees may have different structures by using different methodologies, because of the variations in the imputed values of ETHGRP and BMIB, and different approaches to selecting splits and stopping rules in RPART and PARTY packages, the derived models were consistent in terms of the most predictive covariates of the response variable. In particular, the most predictive covariates of SPINEZ_FIRST are REGION, BISPHOS, SEX and BIWEEK_UP_FIRST, the most predictive covariates of FIRST_RATE_SPINEZ are REGION, SPINEZ_FIRST, and ETHGRP, and the most predictive covariate of RATE_SPINEZ is PREVIOUS_SPINEZ.

The modeling results of SPINEZ_FIRST suggest that the longer the duration between a patient's first infusion of imiglucerase and his/her first DXA assessment visit is, the larger the patient's bone-mineral density tends to be. It can be assumed that most patients had a relatively low bone density at the beginning of the treatment, and it takes some time before the patients can benefit from the ERT treatment. Therefore, if there was a longer period between the start of the treatment and the first DXA assessment, the beneficial effect of the treatment could manifest itself for a longer period, and the bone-mineral density would be higher as a result.

For those who had the treatment with bisphosphonates, female patients tended to have higher initial bone-mineral density than male patients.

Regarding the rate of change in a patient's DXA Z-score between the patient's first infusion and each of his/her following DXA assessment visits, the models suggest that the patients with a lower value of bone-mineral density at the beginning of the treatment caught up to the normal level of bone-density more quickly. It is also interesting to notice that among the patients whose initial DXA Z-score values were from -2.66 to 0.36, those from Europe profited less from the treatment than the patients from other regions.

When the rate of bone-mineral density change between successive measurements is evaluated, the most predictive variable is the previous bone-mineral density value. In particular, patients with lower previous bone-mineral density had more improvement from the treatment. This is again consistent with the theory that patients with lower bone-mineral density have more of a need to catch up.

Another interesting finding is that the treatment dosage only appears twice in the models of FIRST_SPINEZ, and it only affects Arab, Hispanic, Caucasian, Jewish-Ashkenazi, Jewish-Sephardic, and American Indian patients. Among these patients, a higher dosage led to faster improvement of bone-mineral density in general.

None of the new variables constructed from patient's hematological, visceral, and bone manifestations show up as important predictive covariates of the three response variables studied in the research.

List of References:

Allison, D.P. (2001). *Missing data (Quantitative Applications in the Social Sciences)*. Thousand oaks, California: Sage Publications, Inc.

Andersson, H., Kaplan, P., Kacena, K., & Yee, J. (2008). Eight-Year Clinical Outcomes of Long-Term Enzyme Replacement Therapy for 884 Children with Gaucher Disease Type 1. *Journal of Pediatrics*, *122*, 2007-2144.

Bembi, B., Ciana, G., Mengel, E., Terk, M.R., Martini, C., & Wenstrup, R. J. (2002). Bone complications in children with Gaucher disease. *The British Journal of Radiology, 75,* A37–A43.

Bolker, M., Brooks, E., Clark, J., Geange, W., Poulsen, R., Stevens, H., & White, S. (2008). Generalized Linear Mixed Models: A Practical Guide for Ecology and Evolution. *Journal of Trends in Ecology and Evolution, 24,* 127-135.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth, CA: Pacific Grove.

Charrow, J., Dulissec, B., Grabowski, G.A., & Weinreb, N.J. (2007). The effect of enzyme replacement therapy on bone crisis and bone pain in patients with type 1 Gaucher disease. *Journal of Clin Genet*, 71, 205–211.

Chen, G., & Astebro, T. (2003). How to deal with missing categorical data: Test of a simple Bayesian method. *Journal of Organizational Research Methods, 6(3),* 309-327.

Data mining tools. (n.d.). *An Introduction to Data Mining.* Retrieved September 3, 2011, from http://dataminingtools.net/wiki/introduction_to_data_mining.php

Dobra, A. (2002). *Classification and Regression Tree Construction*. Retrieved September 18, 2011, from Cornell University, Department of Computer Science.

Elstein, D., Abrahamov, A., Halpern, I. H., Meyer, A., & Zimran, A. (1998). Low-dose low-frequency imiglucerase as a starting regimen of enzyme replacement therapy for patients with type I Gaucher disease. *Journal of Q J Med*, *91*, 483-488.

Fayyad, M.U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: an Overview. *Advances in Knowledge Discovery and Data Mining* (pp. 1-36). Menlo Park, CA, USA: American Association for Artificial Intelligence.

Journal of American Society of Hemotology. (March 9, 2006) *Superior effects of high-dose enzyme replacement therapy in type 1 Gaucher disease on bone marrow involvement and chitotriosidase levels: a 2-center retrospective analysis*. Retrieved August 11, 2011 from bloodjournal.hematologylibrary.org.

Han, J., Kamber, M., & Pei, J. (2011). Classification: Basic Concepts. *Data Miming: Concepts and Techniques* (pp. 344). CA: AAAI Press.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework, *Journal of Computational and Graphical Statistics, 15,* 651-674.

Hui, S.L., Gao, S.J., Zhou, X.H., Johnston, C.C., Lu, Y., Gluer, C.C., et al. (1997).Universal Standardization of Bone Density Measurements: A Method with Optimal Properties for Calibration Among Several Instruments. *Journal of Bone and Mineral Research*, *12*, 1463-1470.

Kalton, G. & Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Journal of Survey Methodology, 12*, 1-16.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of Applied Statistics, 29*, 119-127.

Little, A. & Rubin, B. D. (1987*). Statistical Analysis with Missing Data*. (2nd e.d.). Hoboken, New Jersey: John & Wiley.

Liu, Y. (2006). *Interactive visual data mining modeling to enhance understanding and effectiveness of the process*. Retrieved September 18, 2011, from Purdue University.

McLachlan, J.G., Do, K.A., & Ambroise, C. (2004). *Analyzing Microarray Gene Expression Data (Wiley Series in Probability and Statistics).* Hoboken, New Jersey: John & Wiley.

Mistry, P.K., Weinreb, N.J., Kaplan, P., Alexander, J.C., Gwosdow, A.R., & Hangartner, T. (2011), Osteopenia in Gaucher disease develops early in life: Response to imiglucerase enzyme therapy in children, adolescents and adults, *Journal of Blood Cells Mol Disease, 46(1),* 66-72.

Murthy, S. K. (1998). Automatic construction of decision trees from data: a multi-disciplinary survey. *Journal of Data Mining and Knowledge Discovery*, 2(4), 345-389.

Poll, L.M., Maas, M., Terk, M.R., Roca-Espiau, M., Bembi, B., Ciana, G., & Weinreb, N. J. (2002). Response of Gaucher bone disease to enzyme replacement therapy. *The British Journal of Radiology, 75,* A25–A36.

Quinlan, J.R. (1986). Induction of Decision Trees. *Journal of Machine Learning, 1*, 81-106.

Quinlan, J.R. (1993). *C 4.5 -- Programs for Machine Learning*. CA: Morgan Kaufmann.

Rosenthal, D.I., Doppelt, S.H., Mankin, H.J., Dambrosia, J.M., Xavier, R.J., McKusick, K.A., et al. (1995). Enzyme replacement therapy for Gaucher

disease: skeletal responses to macrophage-targeted glucocerebrosidase. *Journal of Pediatrics, 96,* 629–637.

Roth, P. (1994). Missing data: A conceptual review for applied psychologists. *Journal of Personnel Psychology, 47,* 537-560.

Roth, P. L., & Switzer, F. S. (1995). A Monte Carlo Analysis of Missing Data Techniques in a HRM Setting. *Journal of Management, 21(5),* 1003-1023.

Sims, K.B., Pastores, G.M., Weinreb, N.J., Barranger, J., Rosenbloom, B.E., Packman, S., et al. (2008). Improvement of bone disease by imiglucerase (Cerezyme) therapy in patients with skeletal manifestations of type 1 Gaucher disease: results of a 48-month longitudinal cohort study. *Journal of Clin Genet, 73,* 430–440.

The Data Mining Process. (n.d.) Retrieved September 3, 2011, from http://www.dataminingexpertsolutions.com/dm-process/

Therneau, M., Atkinson, M., Foundation, M. (1997). *An Introduction to Recursive Partitioning Using the RPART Routines.* Retrieved May 18, 2011, from http://www.mayo.edu/hsr/techrpt/61.pdf

Timofeev, R. (2004). *Classification and Regression Trees (CART) Theory and Applications*. Retrieved December 6, 2011, from Center of Applied Statistics and Economics Humboldt University, Berlin.

Tóth, J., Szücs F.Z., Benkö, K., Maródi, L. (2003). Enzyme replacement therapy in Gaucher disease: monitoring visceral and bone changes with MRI. *Journal of Orv Hetil, 144,* 749-755.

Wayman. J. (2003). *Multiple Imputation For Missing Data: What Is It And How Can I Use It?* Retrieved May 11, 2011, from Center for Social Organization of Schools, John Hopkins University.

Weinreb, N. J., Charrow, J., Andersson, H. C., Kaplan, P., Kolodny, E. H., Mistry, P., et al. (2002). Effectiveness of enzyme replacement therapy in 1028 patients with type 1 Gaucher disease after 2 to 5 years of treatment: a report from the Gaucher Registry, *Journal of Excerpta Medica, 113,* 112-119.

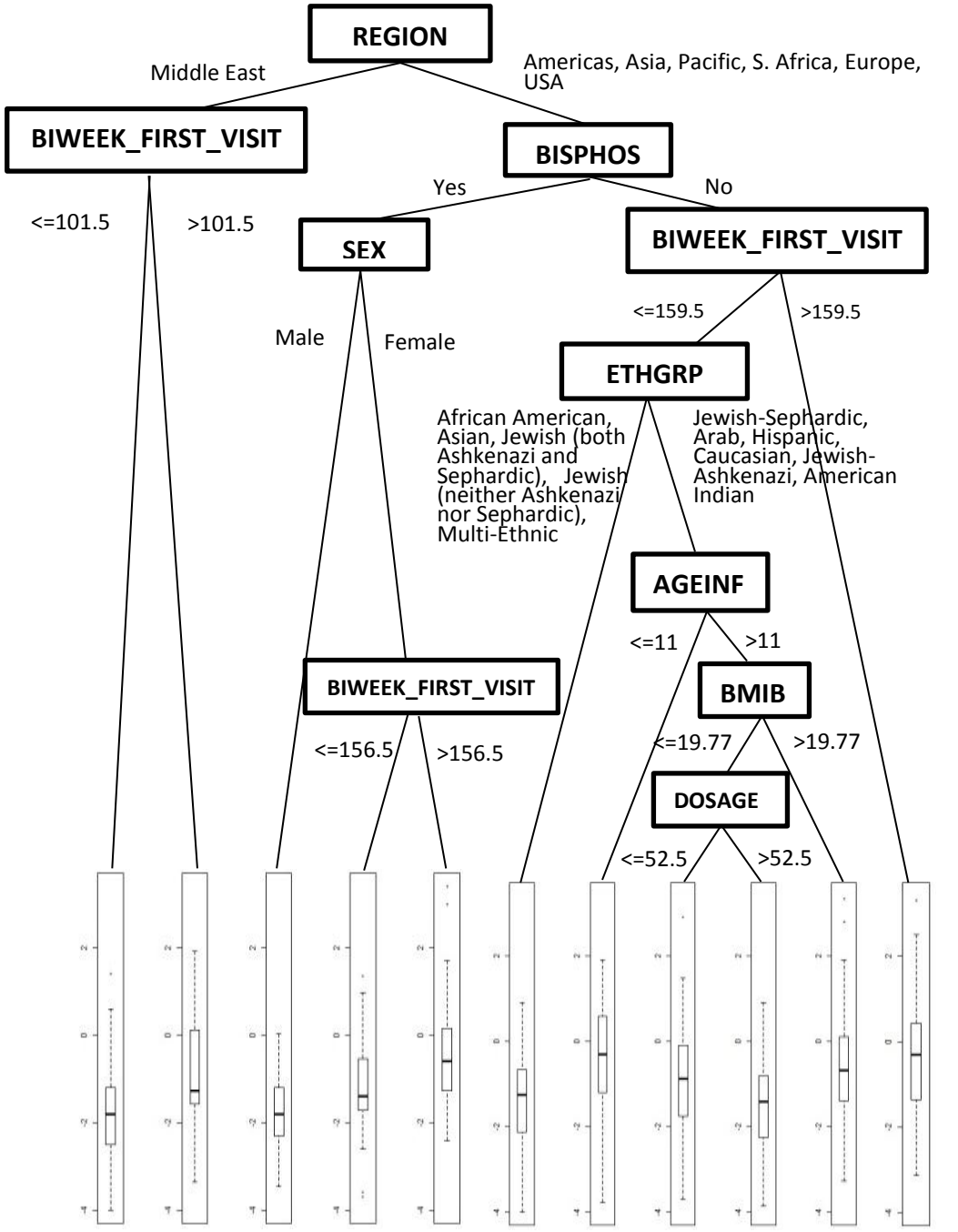
Weinreb, N., Barranger, J., Packman, S., Prakash-Cheng, A., Rosenbloom, B., Sims, K., et al. (2007). Imiglucerase (Cerezyme) improves quality of life in patients with skeletal manifestations of Gaucher disease. *Journal of Clinical Genetics, 71,* 576-588.

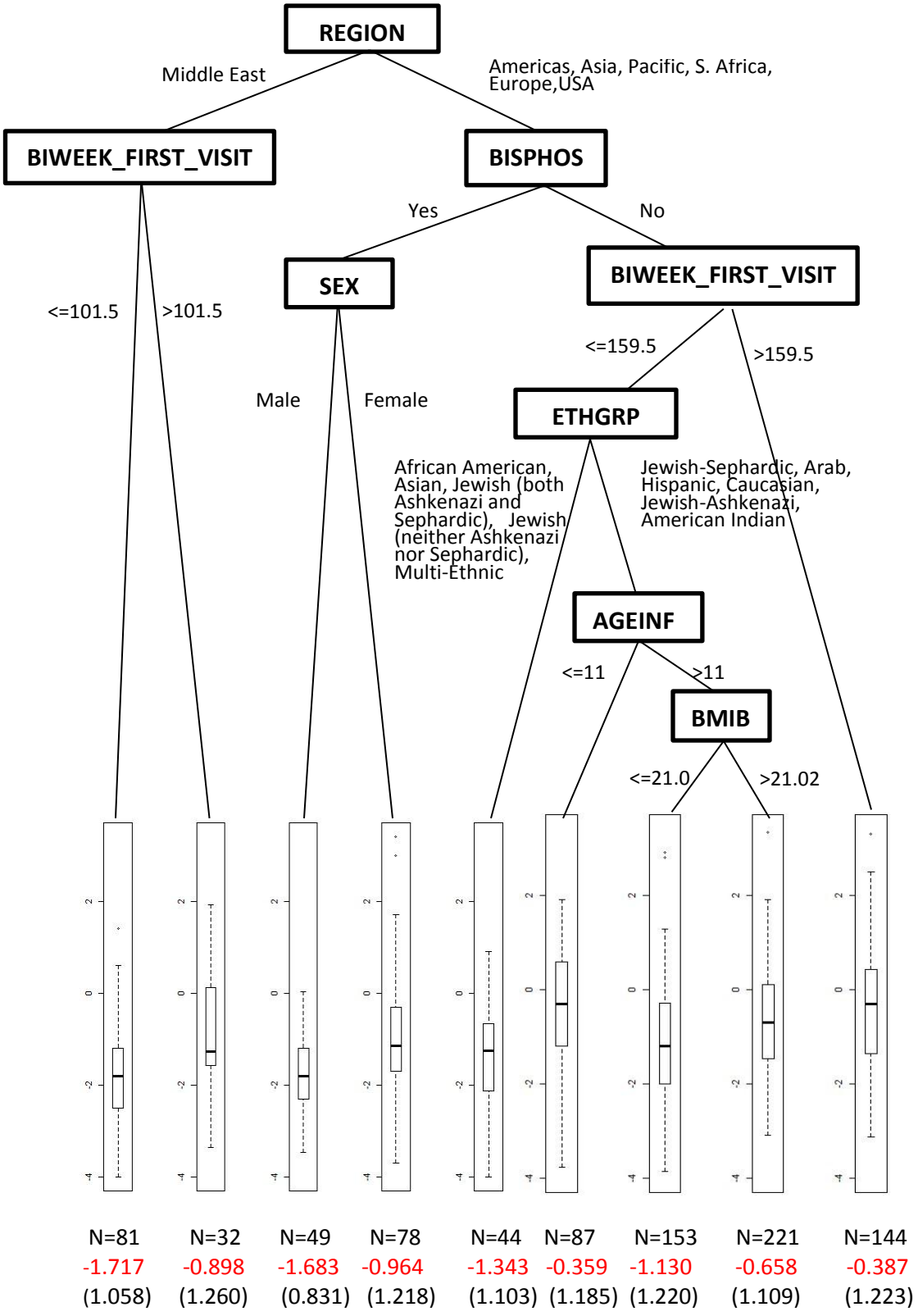
Wenstrup, R.J., Kacena, K.A., Kaplan, P., Pastores, G.M., Cheng, A.P., Zimran, A., et al (2007). Effect of Enzyme Replacement Therapy with Imiglucerase on BMD in Type 1 Gaucher Disease. *Journal of Bone and Mineral Research, 22.*

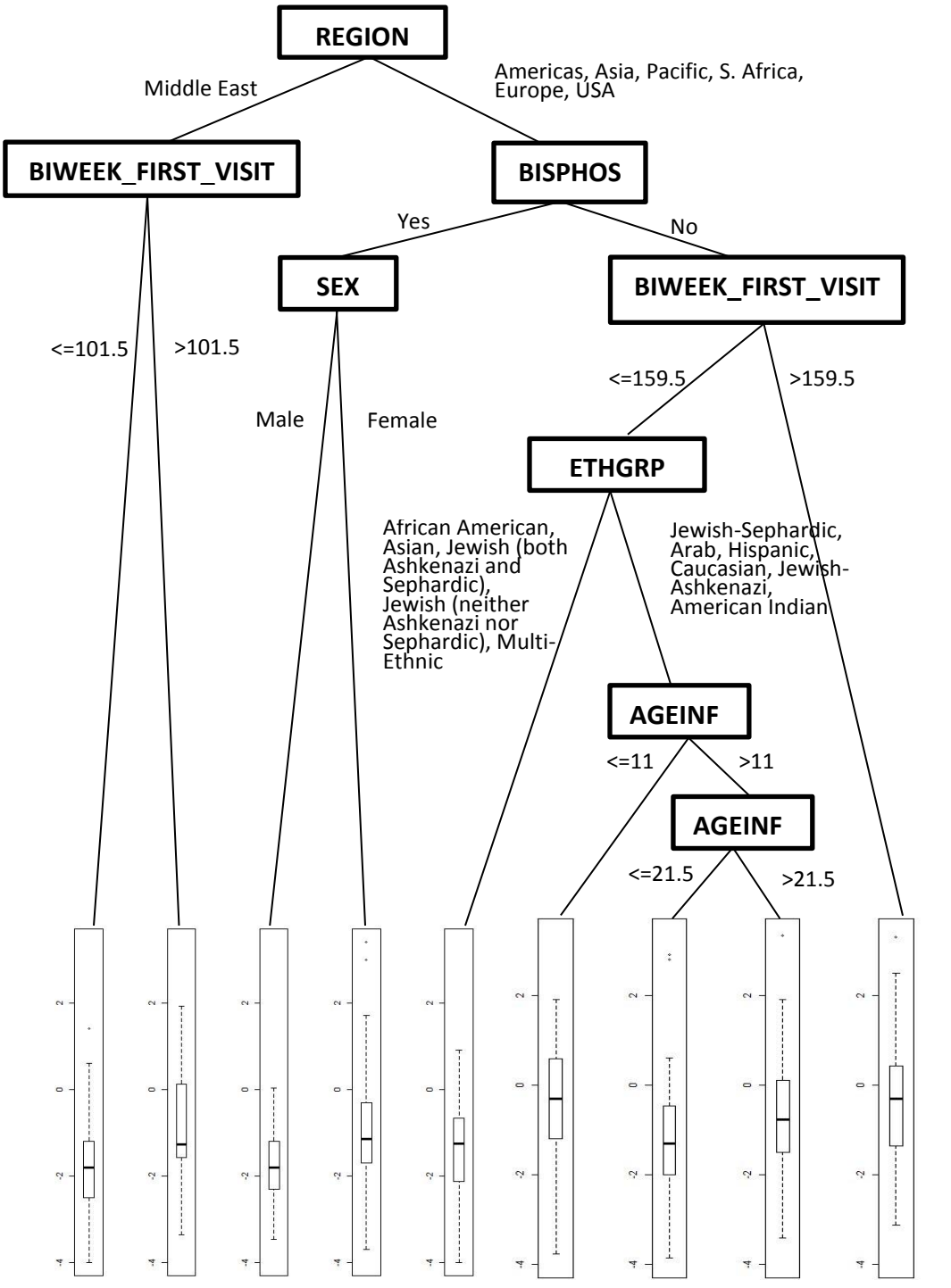
West, B., Welch, K., & Galecki, A. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Boca Raton: Chapman & Hall/CRC.

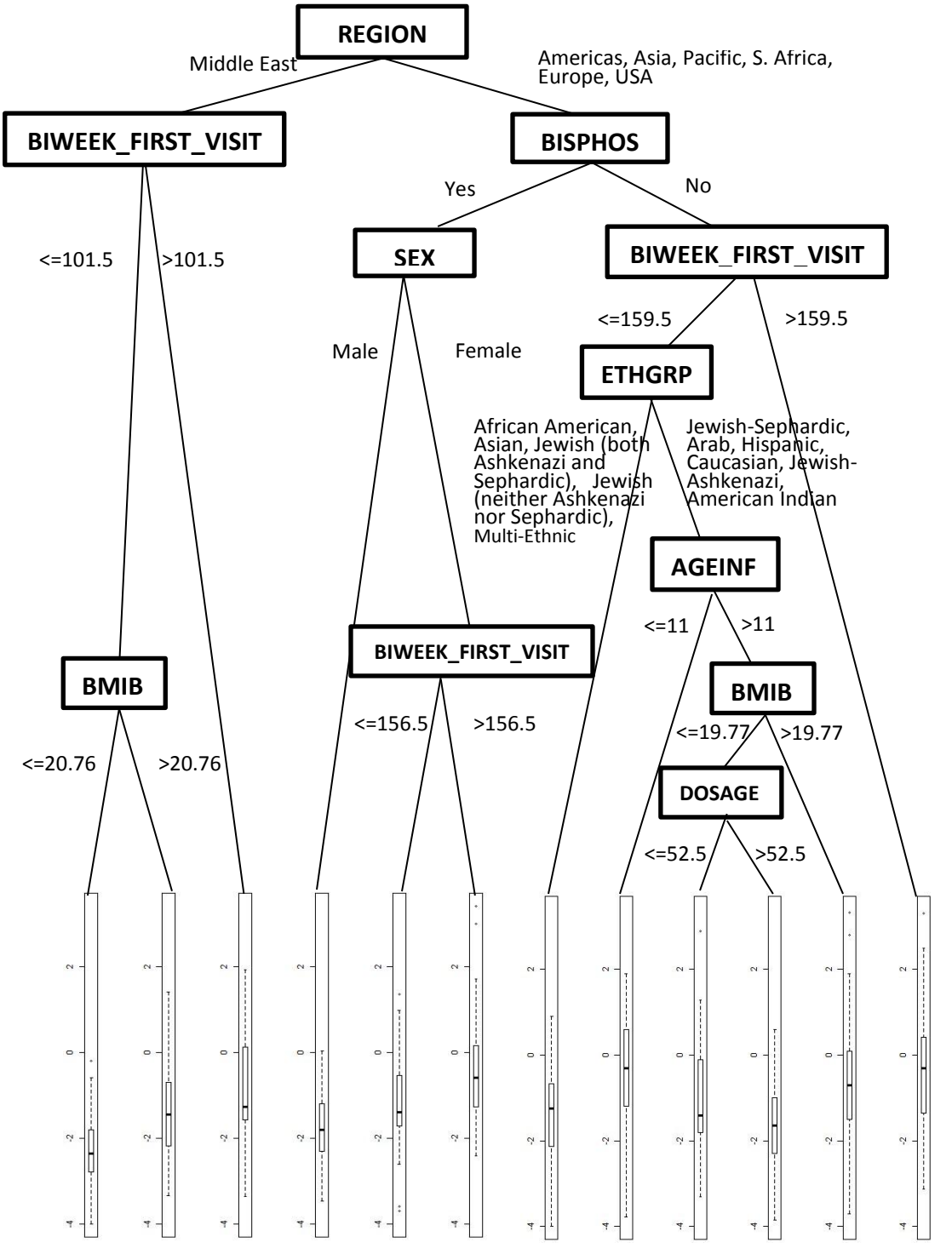
APPENDIX. Models of SPINEZ_FIRST

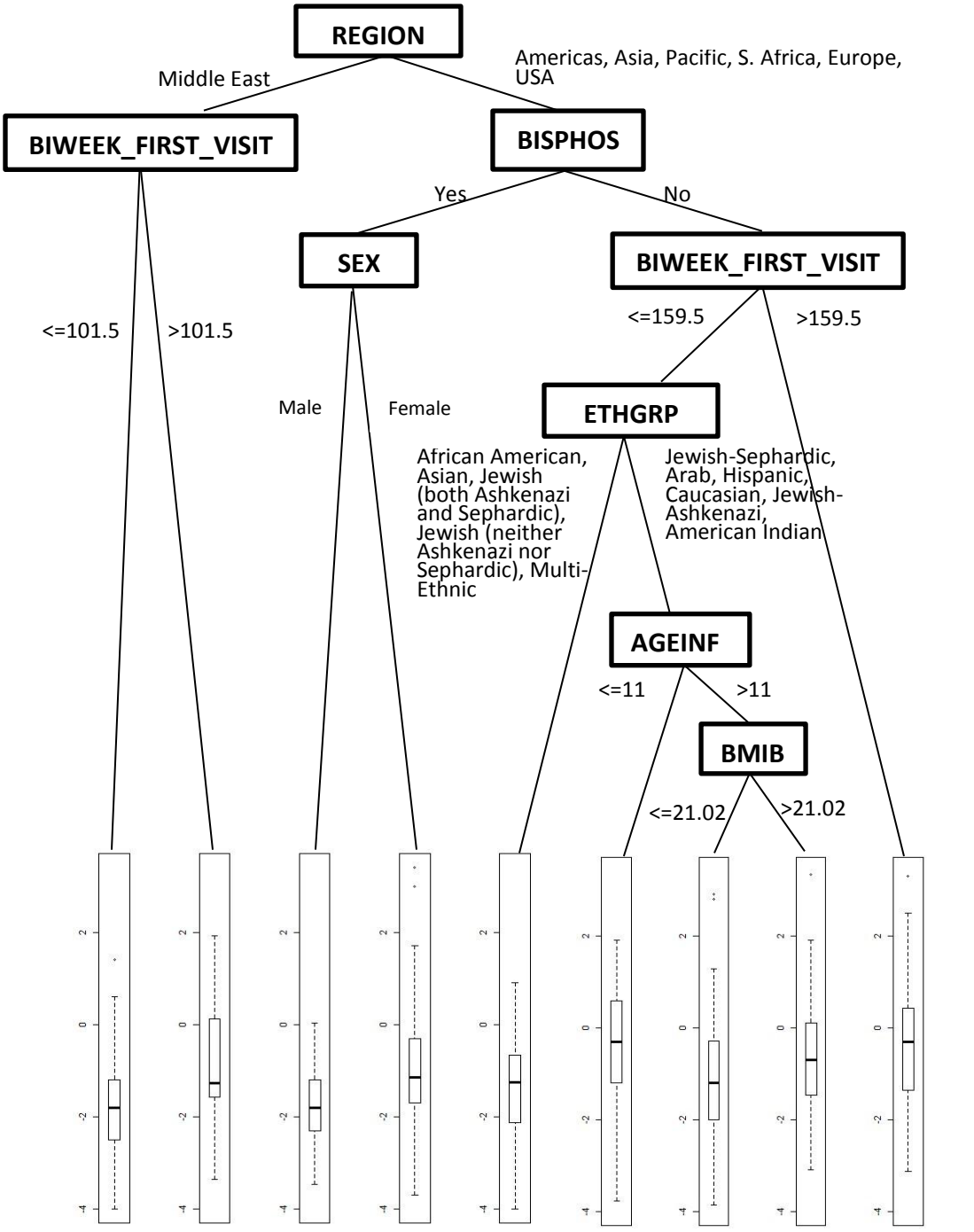
This appendix contains 9 models of SPINEZ_FIRST derived using the RPART package from complete datasets, which are not concluded in 4.2.2 of the report.

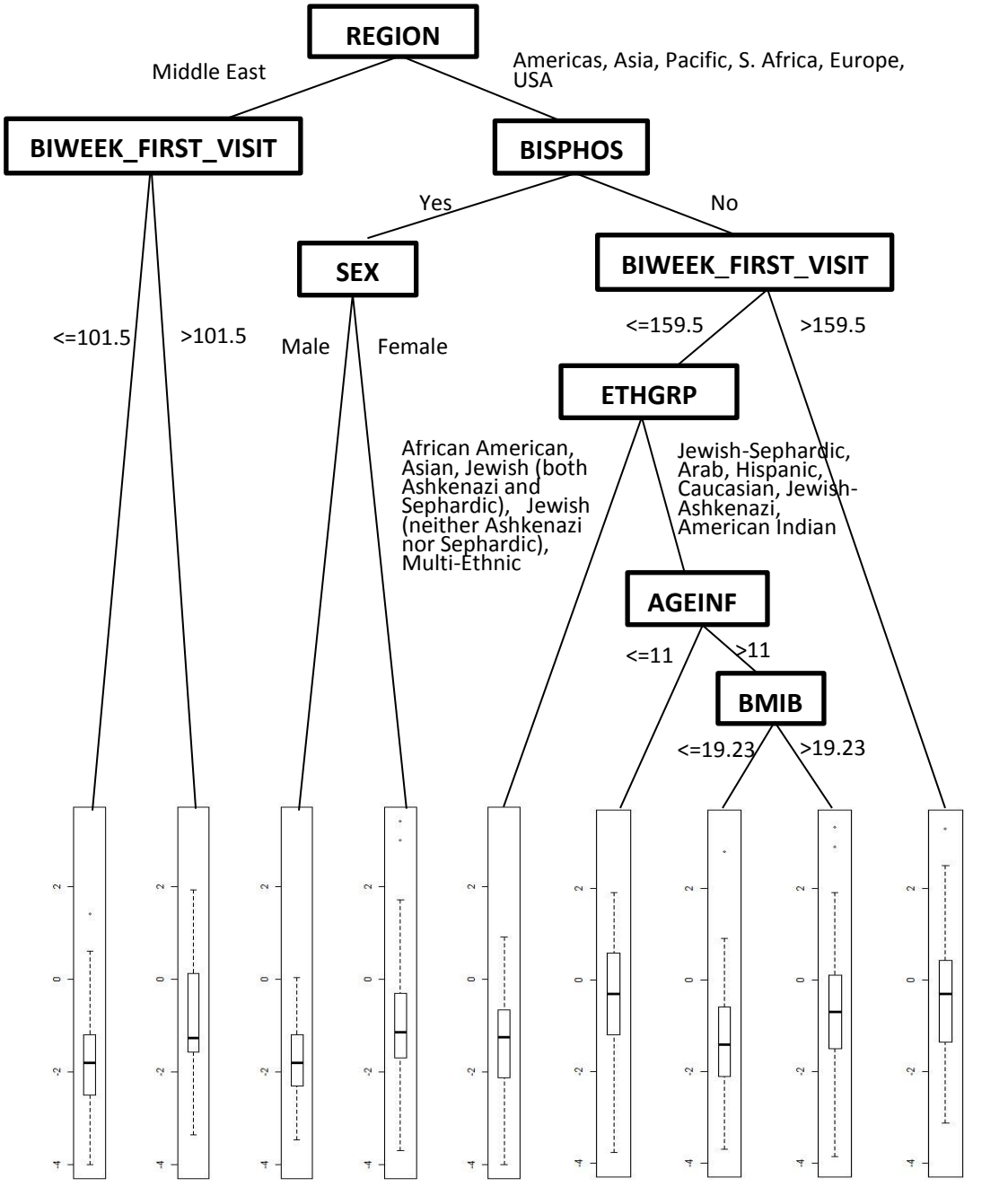


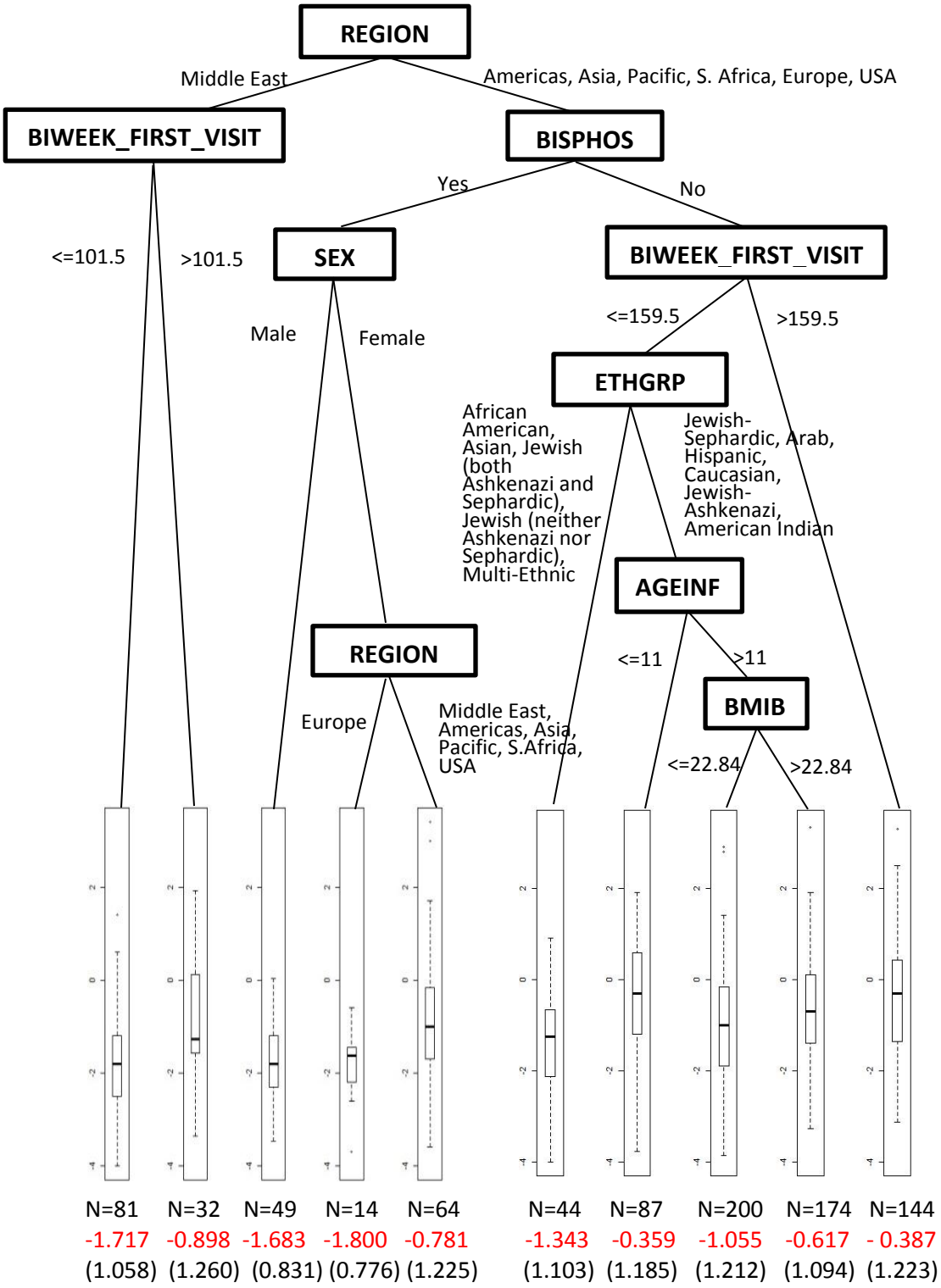

REGION
Middle East
Americas, Asia, Pacific, S. Africa, Europe,USA
BIWEEK_FIRST_VISIT
BISPHOS
Yes
No
SEX
BIWEEK_FIRST_VISIT
<=101.5
>101.5
<=159.5
>159.5
Male
Female
ETHGRP
African American, Asian, Jewish (both Ashkenazi and Sephardic), Jewish (neither Ashkenazi nor Sephardic), Multi-Ethnic
Jewish-Sephardic, Arab, Hispanic, Caucasian, Jewish-Ashkenazi, American Indian
AGEINF
<=11
>11
BMIB
<=21.0
>21.02
N=81 -1.717 (1.058)
N=32 -0.898 (1.260)
N=49 -1.683 (0.831)
N=78 -0.964 (1.218)
N=44 -1.343 (1.103)
N=87 -0.359 (1.185)
N=153 -1.130 (1.220)
N=221 -0.658 (1.109)
N=144 -0.387 (1.223)

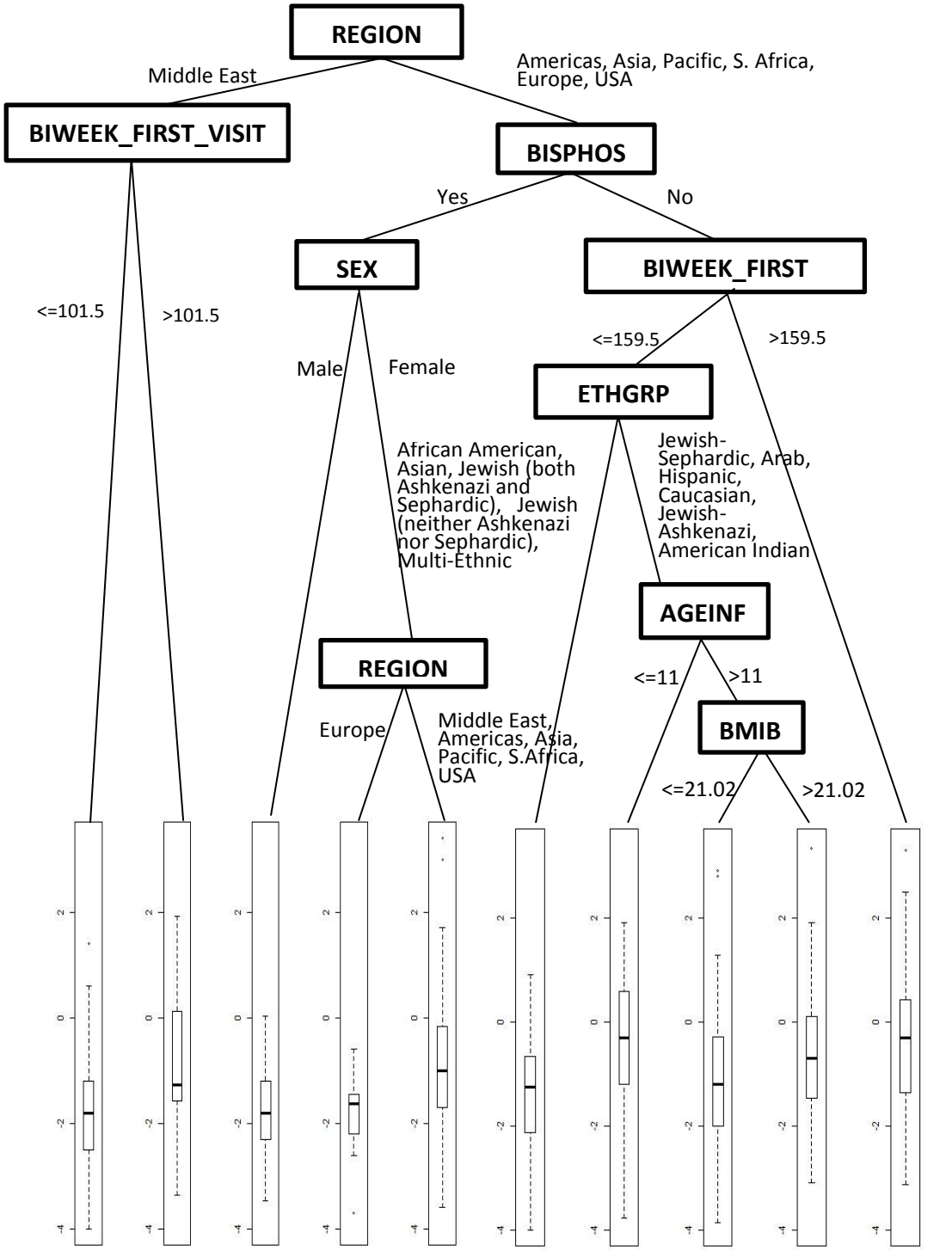

REGION
Middle East
Americas, Asia, Pacific, S. Africa, Europe, USA
BIWEEK_FIRST_VISIT
BISPHOS
Yes
No
SEX
BIWEEK_FIRST_VISIT
<=101.5
>101.5
Male
Female
<=159.5
>159.5
ETHGRP
African American, Asian, Jewish (both Ashkenazi and Sephardic), Jewish (neither Ashkenazi nor Sephardic), Multi-Ethnic
Jewish-Sephardic, Arab, Hispanic, Caucasian, Jewish-Ashkenazi, American Indian
AGEINF
<=11
>11
AGEINF
<=21.5
>21.5
N=81 N=32 N=49 N=78 N=44 N=87 N=91 N=283 N=144
-1.717 -0.898 -1.683 -0.964 -1.343 -0.359 -1.233 -0.728 -0.387
(1.058) (1.260) (0.831) (1.218) (1.103) (1.185) (1.207) (1.143) (1.223)

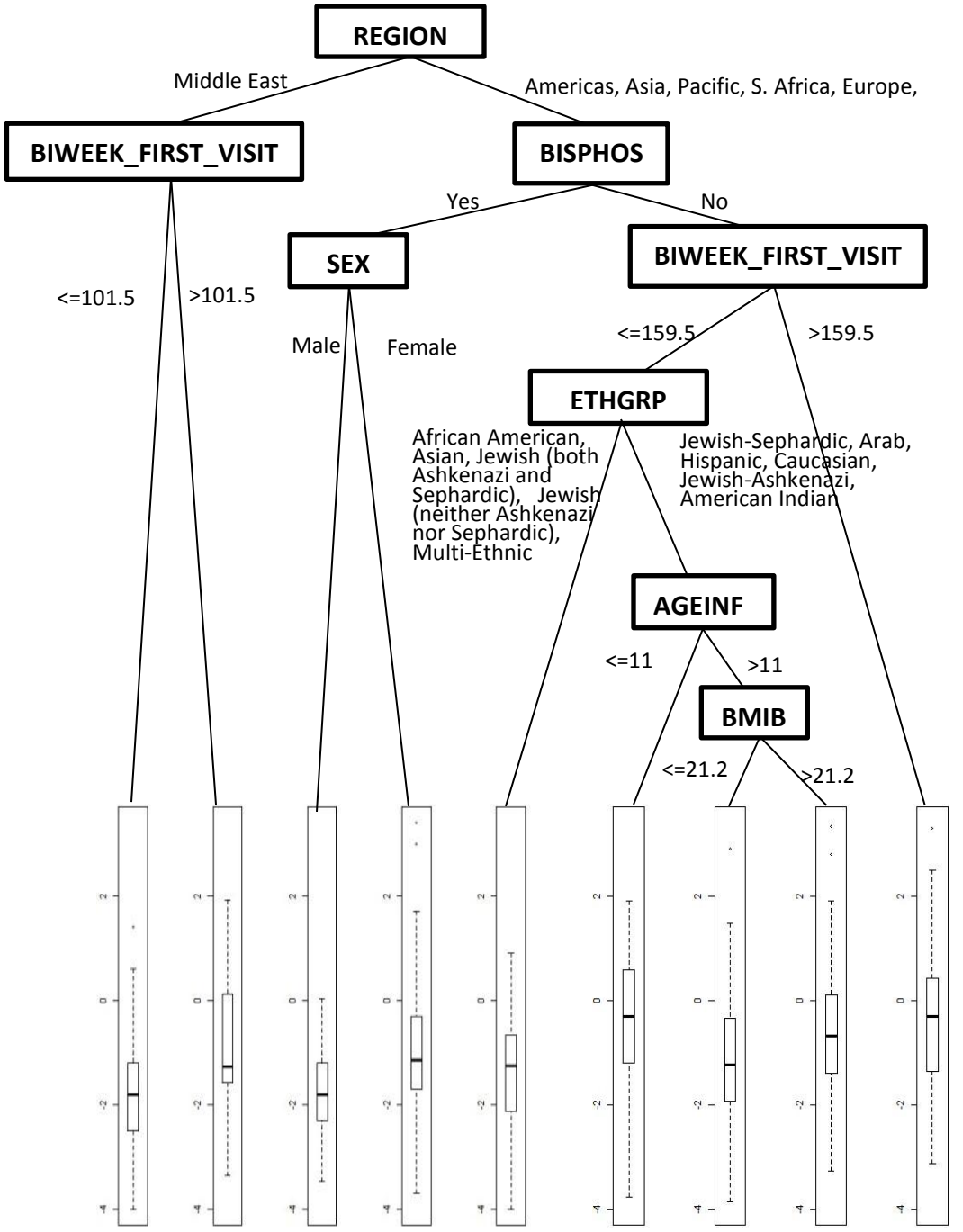

REGION
Middle East
Americas, Asia, Pacific, S. Africa, Europe, USA
BIWEEK_FIRST_VISIT
BISPHOS
<=101.5
>101.5
Yes
No
SEX
BIWEEK_FIRST_VISIT
Male
Female
<=159.5
>159.5
ETHGRP
African American, Asian, Jewish (both Ashkenazi and Sephardic), Jewish (neither Ashkenazi nor Sephardic), Multi-Ethnic
Jewish-Sephardic, Arab, Hispanic, Caucasian, Jewish-Ashkenazi, American Indian
AGEINF
<=11
>11
BIWEEK_FIRST_VISIT
BMIB
<=156.5
>156.5
BMIB
<=19.77
>19.77
<=20.76
>20.76
DOSAGE
<=52.5
>52.5
N=30 N=51 N=32 N=49 N=58 N=20 N=44 N=87 N=58 N=38 N=278 N=144
-2.291 -1.380 -0.898 -1.683 -1.189 -0.313 -1.343 -0.359 -0.996 -1.626 -0.715 -0.387
(0.806) (1.050) (1.260) (0.831) (1.020) (1.51) (1.103) (1.185) (1.221) (1.084) (1.139) (1.223)


REGION
Middle East
Americas, Asia, Pacific, S. Africa, Europe, USA
BIWEEK_FIRST_VISIT
BISPHOS
Yes
No
SEX
BIWEEK_FIRST_VISIT
<=101.5
>101.5
<=159.5
>159.5
Male
Female
ETHGRP
African American, Asian, Jewish (both Ashkenazi and Sephardic), Jewish (neither Ashkenazi nor Sephardic), Multi-Ethnic
Jewish-Sephardic, Arab, Hispanic, Caucasian, Jewish-Ashkenazi, American Indian
AGEINF
<=11
>11
BMIB
<=21.02
>21.02
N=81 N=32 N=49 N=78 N=44 N=87 N=153 N=221 N=144
-1.717 -0.898 -1.683 -0.964 -1.343 -0.359 -1.130 -0.658 -0.387
(1.058) (1.260) (0.831) (1.218) (1.103) (1.185) (1.220) (1.109) (1.223)


REGION
Middle East
Americas, Asia, Pacific, S. Africa, Europe, USA
BIWEEK_FIRST_VISIT
BISPHOS
Yes
No
SEX
BIWEEK_FIRST_VISIT
<=101.5
>101.5
Male
Female
<=159.5
>159.5
ETHGRP
African American, Asian, Jewish (both Ashkenazi and Sephardic), Jewish (neither Ashkenazi nor Sephardic), Multi-Ethnic
Jewish-Sephardic, Arab, Hispanic, Caucasian, Jewish-Ashkenazi, American Indian
AGEINF
<=11
>11
BMIB
<=19.23
>19.23
N=81 N=32 N=49 N=78 N=44 N=87 N=89 N=285 N=144
-1.717 -0.898 -1.683 -0.964 -1.343 -0.359 -1.314 -0.706 -0.387
(1.058) (1.260) (0.831) (1.218) (1.103) (1.185) (1.151) (1.150) (1.223)


REGION
Middle East
Americas, Asia, Pacific, S. Africa, Europe, USA
BIWEEK_FIRST_VISIT
BISPHOS
Yes
No
<=101.5
>101.5
SEX
BIWEEK_FIRST_VISIT
Male
Female
<=159.5
>159.5
ETHGRP
African American, Asian, Jewish (both Ashkenazi and Sephardic), Jewish (neither Ashkenazi nor Sephardic), Multi-Ethnic
Jewish-Sephardic, Arab, Hispanic, Caucasian, Jewish-Ashkenazi, American Indian
AGEINF
REGION
<=11
>11
BMIB
Europe
Middle East, Americas, Asia, Pacific, S.Africa, USA
<=22.84
>22.84
N=81 N=32 N=49 N=14 N=64 N=44 N=87 N=200 N=174 N=144
-1.717 -0.898 -1.683 -1.800 -0.781 -1.343 -0.359 -1.055 -0.617 - 0.387
(1.058) (1.260) (0.831) (0.776) (1.225) (1.103) (1.185) (1.212) (1.094) (1.223)


REGION
Middle East
Americas, Asia, Pacific, S. Africa, Europe, USA
BIWEEK_FIRST_VISIT
BISPHOS
Yes
No
SEX
BIWEEK_FIRST
<=101.5
>101.5
Male
Female
<=159.5
>159.5
ETHGRP
African American, Asian, Jewish (both Ashkenazi and Sephardic), Jewish (neither Ashkenazi nor Sephardic), Multi-Ethnic
Jewish-Sephardic, Arab, Hispanic, Caucasian, Jewish-Ashkenazi, American Indian
AGEINF
REGION
<=11
>11
Europe
Middle East, Americas, Asia, Pacific, S.Africa, USA
BMIB
<=21.02
>21.02
N=81 N=32 N=49 N=14 N=64 N=44 N=87 N=153 N=216 N=144
-1.717 -0.898 -1.683 -1.800 -0.781 -1.343 -0.359 -1.130 -0.657 -0.387
(1.058) (1.260) (0.831) (0.776) (1.225) (1.103) (1.185) (1.220) (1.120) (1.223)



N=81	N=32	N=49	N=78	N=44	N=87	N=168	N=206	N=144
-1.717	-0.898	-1.683	-0.964	-1.343	-0.359	-1.128	-0.625	-0.387
(1.058)	(1.260)	(0.831)	(1.218)	(1.103)	(1.185)	(1.163)	(1.142)	(1.223)